교육과정평가연구
The Journal of Curriculum & Evaluation
2006, Vol. 9, No. 2, pp. 121～138

개인화 종합평가시스템 탐색

성 태 제
(이화여자대학교 교수)

≪ 요 약 ≫

교육평가의 기능은 개인의 성장·발달을 촉진시키는 것이다. 그러므로 개인 위주의 평가체제로 변화하여야 하고, 이런 관점에서 능력과 성장참조평가를 실시하여야 하며, 개인의 능력수준에 맞는 검사를 실시하여야 한다. 고등정신능력 함양을 위하여 수행평가가 선호되고 다양한 방법에 의한 검사와 측정결과에 의하여 개인을 평가하는 총평체제로 변화되며, 일회적 평가에서 지속적 평가를 위하여 보다 과학적인 문항반응이론이 적용하여야 한다. 인간에 대한 평가는 종합적이어야 하기에 새로운 평가체제를 적용한 개인화 종합평가시스템을 제안한다. 개인화 종합평가시스템은 컴퓨터를 이용하여, 개인 능력수준에 맞는 평가도구를 제시하고 정답 여부에 따라 다른 문제가 제시된다. 검사나 측정 결과를 컴퓨터에 저장하여 지속적으로 개인의 변화·성장을 분석하여 평가를 실시하며, 면접이나 관찰결과도 종합하여 개인을 총평할 수 있다. 개인화 종합평가시스템을 개발하기 위하여 교육평가전문가, 교과내용전문가, 그리고 컴퓨터 관련 전문가들의 협동 연구가 필요하고 이를 위하여 국가적으로 인적자원 개발 차원에서의 지원 아래 공동 개발 작업이 요구된다.

주제어 : 성장참조평가, 컴퓨터화 검사, 수행평가, 총평, 종합평가

Ⅰ. 서론

사람 '人'자는 ノ자와 丶자로 합성되어 있다. 각기 다른 두 글자가 서로를 받치고 있어 넘어지지 않게 균형을 이루고 있다. 그러므로 사람은 각자를 독특한 개체로 존중하고 서로 도우며 존재한다. 사람을 대상으로 하는 교육은 '가르치다'와 '기르다'의 합성어로 '갈고 닦게 하는 훈련을 통하여 사람을 성장시킨다'를 의미한다. Pedagogy는 그리스어의 paidagogos에서 유래되며, 아동(paidos)과 이끌다(agogos)의 합성어로 '미성숙학 아동을 가르치고 양육한다'의 의미를 지니고 있으며 Education은 라틴어 educare와 educere에서 유래된 용어로 '이끌어 내다'라는 의미를 지닌다(유재봉, 2007). 정범모(1968)는 교육의 개념을 '인간행동의 계획적인

변화'라 하였고, 이 정의에 의하면 교육의 근본 목적은 인간의 행동을 변화시키는 것이다. 인간의 행동은 두뇌와 관련된 '인지적 영역', 가슴과 관련된 '정의적 영역', 신체 활동과 관련된 '심동적 영역'으로 구분된다.

교육평가란 교육의 행위를 극대화시키기 위한 교육의 보조수단으로 교육과 관련된 모든 것에 대한 가치판단이다. 평가대상에는 학생들의 학업성취도, 가치관, 태도, 교사의 능력, 학교시설, 재정, 교육의 주변 환경 등 교육과 관련된 모든 것이 포함된다. 교육의 극대화를 위해 교육과정을 개정하고, 새로운 교수법을 도입하며, 학급당 학생 수를 포함하여 교육 여건을 개선시키고 교사의 질을 향상시키기 위한 노력들이 이루어지고 있으며, 이 과정에서 평가가 중요한 역할을 하고 있다. 최근 학교교육의 책무성을 강화하기 위해 학업성취도 평가 및 학교평가를 지속적으로 실시하고 있으며, 특정 분야의 능력수준을 측정하기 위한 인증시험이 다양한 분야에서 시행되고 있고, 컴퓨터와 인터넷의 발달로 문제은행의 구축을 통한 컴퓨터화 검사의 시행이 확대되고 있는 추세이다.

이러한 변화 중에서도 주목할 점은 인간 개인을 독특한 개체로 인정하고 개인의 변화와 발전을 위한 평가체제로 변모하고 있다는 점이다(성태제, 2001; Osterhof, 2001). 교육현장에서도 선택형 중심의 집단적 대단위 평가보다는 학생 개인의 다양한 특성을 평가하기 위한 수행평가가 실시되고 있으며, 검사에 있어서도 지필검사보다는 컴퓨터화 검사로, 컴퓨터화 검사 중에서도 학생의 능력수준에 따라 다른 문항이 제시되는 컴퓨터화 능력적응검사로 변화되고 있다. 학생 선발에 있어서도 특정 영역에서 우수한 학생을 선발하는 기회를 확대하고 있다.

이러한 변화는 예전의 획일적이고 기계적이며 집단적 방법에 의한 교육이나 교육평가 방법이 교육의 목적에 부합하기 어려울 뿐 아니라 개인의 성장·발달이나 사회와 국가발전에 도움이 되지 않는다는 경험적 판단에서 비롯된 것이라 할 수 있다. 검사만으로 인간을 평가하던 고전적 방법에서 검사 이외의 다른 방법들을 활용하여 다양한 측면을 평가하는 총평(assessment)체제로 변하고 있으며, 한 번의 평가로 끝내는 것이 아니라 일정시간 간격을 설정하여 지속적으로 개인을 평가하여 성장과 변화를 분석하는 평가형태로 발전하고 있다. 뿐만 아니라 문제해결의 장애 원인을 밝히고 교정하여 성장과 발달을 촉진하는 지적 측정(intelligent measurement)을 실시하여 교육평가의 본질을 추구하고 있다.

교육평가의 특성은 과학적이며, 체계적이고, 지속적이며, 종합적이어야 한다는 것이다. 이에 비추어 볼 때 현재까지의 교육평가는 고전적 평가방법에 의하여 일시적으로 진행되어 왔으며, 개인에 대한 정보에 대하여 부분적으로 분석하여 평가하여 온 경향이 높다. 개인에 대한 평가는 측정과 검사보다는 총평을 지향하고 있으며, 지속적인 변화 정도도 평가하는 것이 개인을 이해하는 데 바람직하다. 본 연구는 교육평가의 특징을 적용하며, 교육평가의 본질에 따른 교육평가 방법의 변화와 개인을 중심으로 한 종합평가시스템을 탐색하고 우리나라 교육의 장래를 위한 적용방법을 논의하고자 한다.

Ⅱ. 교육평가 방법의 변화

1. 상대비교평가에서 성장참조평가로

인간의 능력을 평가하던 고전적 평가방법은 상대비교평가와 절대평가이다. 상대비교평가는 개인이 획득한 점수의 상대적 서열에 의하여 평가하는 방법으로서 상대적 우열의 파악이 용이하나 교수·학습 이론에 적합하지 않은 문제점을 지니고 있다. 절대평가는 개인이 이룬 성취수준과 판단의 기준이 되는 준거(criterion, standard, cut-off, cut score)에 의하여 평가함으로써 준거에 비추어 달성도를 파악하고 pass나 fail의 판정이 용이하다. 상대비교평가는 피험자 집단의 특성을 나타내는 규준에 비추어 평가하므로 규준참조평가(norm-referenced evaluation)라 하고, 절대평가는 준거에 의하여 평가하므로 준거참조평가(criterion-referenced evaluation)라 한다. 두 평가방법은 모두, 평가대상인 개인과, 개인의 특성을 고려하지 않은 점수와 비교한다.

두 평가방법이 지니는 공통적 제한점을 해결하기 위하여 능력참조평가와 성장참조평가가 제안되었다. 인간은 각기 다른 특성을 가지고 태어났고, 추구하는 방향이 서로 다르므로 개개인이 독특한 개체로 인정받을 수 있는 평가방법을 실시하여야 한다는 주장이다.

각기 다른 영역에서 다른 수준의 능력을 지니고 있기에 이를 고려하여 평가하는 것이 보다 인간적이고 교육적이라는 아이디어에서 출발한 이론이 능력참조평가이다. 능력참조평가(ability-referenced evaluation)는 개인이 지니고 있는 능력에 비추어 최대한의 노력(maximum performance)을 하였는가에 주안점을 두는 평가로서 '능력을 최대한 발휘하였는가'와 '충분한 시간이 부여되었을 때 더 잘할 수 있었는가'를 고려한다.

나아가 평가는 지속적으로 이루어지며 개인의 변화에 주안점을 두는 것이 보다 교육적이라는 점에서 성장참조평가가 제안되었다. 황정규(1998)도 개인 수준의 성장을 변화로 표현하고, 변화의 내용과 정도를 측정·평가하는 방법을 갖고 있었다면 교육은 혁신적인 변모를 겪었을 것이라고 주장하고 있다. 성장참조평가(growth-referenced evaluation)란 교육의 진행 과정을 통하여 얼마나 성장하였느냐에 관심을 두는 평가로서 최종적으로 달성한 성취수준보다는 어떻게 얼마만큼의 성장·변화를 가져왔느냐에 관심을 둔다(성태제, 2002; Oosterhof, 1994, 2001).

능력참조평가와 성장참조평가는 비교의 대상이 개인과 본인 그 자체로서 개인을 존중하는 평가방법이라 할 수 있다. 행정적 기능이 강조되는 평가상황에서 두 가지의 평가방법을 적용하는 데는 어려움이 있으나 인간 개인의 행동변화를 추구하는 교육의 근본 목적을 달성하기 위해서는 두 평가방법이 적용되어야 할 것이다.

2. 지필검사에서 컴퓨터화 능력적응검사로

컴퓨터의 발명은 인간 생활의 많은 변화를 가져왔고, 특히 개인 컴퓨터의 발명과 발달은 교육 분야에 엄청난 변화를 가져왔다. 특히 교육평가 분야에서 개인 컴퓨터는 검사의 개념을 변화시키고 발전시켰다. 지필에 의한 시험은 일반적으로 동일한 문항에 다수의 학생들이 응답하는 집단검사의 성격을 지니고 있다. 여기에 두 가지의 문제점이 있다. 하나는 검사윤리 문제이고, 다른 하나는 측정학적 문제이다.

모든 학생에게 동일한 문제가 제시되었을 때 능력이 탁월한 학생은 소홀하거나 부주의로 문항의 답을 틀릴 수 있으며, 쉬운 난이도의 검사 때문에 학습의욕을 잃게 된다. 반대로 문항이 매우 어렵게 느껴지는 학생은 검사 불안으로 실수를 유발하고, 더욱 어려운 시험은 해당 과목에 대한 흥미를 잃어 학습을 포기하므로 끝내는 부정적 자아개념이 형성되는 문제가 나타난다. 이를 검사학대(test abuse)라 한다. 교육에서 검사학대는 교육평가의 본질을 벗어나는 비교육적인 현상으로서 검사 윤리에 위배되는 것이다.

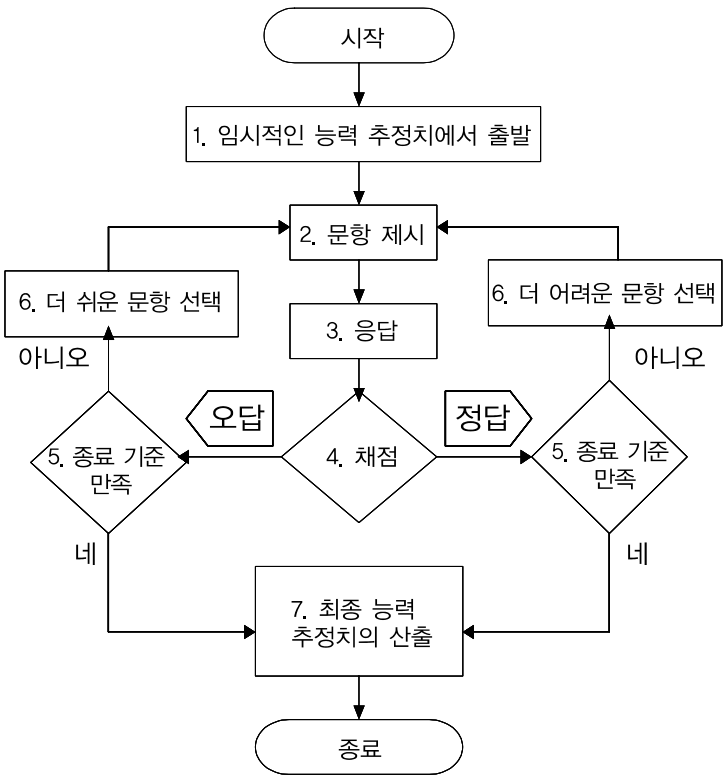
측정학적 관점에서 위에 제시한 두 가지 경우는 측정의 오차를 유발한다. 매우 쉬운 문제를 접한 학생은 부주의나 실수로 문항의 답을 맞히지 못할 것이고, 매우 어려운 문제를 접한 학생은 불안으로 해당 문제 뿐 아니라 다음 문제에 대하여 개인의 능력을 발휘하기가 쉽지 않다. 이로 인하여 생긴 오차는 신뢰도를 저하시켜 측정의 기본목적을 달성하지 못한다.

신뢰도를 저하시키는 측정학적 문제점을 해결하기 위하여 Lord(1970)는 맞춤검사(tailored testing)를 제안하였다. 개인의 몸에 맞는 양복이 있듯이 검사도 각 개인의 능력에 부합하는 문항을 제시하는 것이 측정의 오차를 줄여 신뢰도를 증가시키고 학습동기와 흥미를 유발할 수 있다는 것이다. 이는 Binet와 Simon(1905)이 지능을 측정하기 위하여 문항들을 제작한 후, 연령 수준에 맞는 문항들을 선정하기 위하여 연령별 정답비율에 의하여 문항을 선택하였던 개념의 확장이라 할 수 있다.

맞춤검사는 문항에 대한 응답의 정답 여부에 따라 다른 문항이 제시되는 검사방법으로 지필검사로 실시하는 데 어려움이 있었으나 컴퓨터의 발달로 이 문제를 해결하였다. 특히 개인 컴퓨터가 발전하면서, 컴퓨터를 이용하여 지필검사를 대신하던 컴퓨터 이용검사에서 1990년 이후 맞춤검사의 원리를 이용한 컴퓨터화 능력적응검사(computerized adaptive test)가 실시되고 있다([그림 1] 참조). TOEFL의 Listening Comprehension의 경우, 문항의 답을 맞히면 그 다음의 문항은 다소 어려운 단어와 빠른 속도의 듣기 문항이 제시되고, 틀리면 늦은 속도와 쉬운 내용의 문항이 제시되고 있다.

컴퓨터화 능력적응검사의 측정학적 난제나 컴퓨터의 하드웨어적 장애가 해결되어, 미국에서 널리 이용되고 있다(Bergstorm & Lunz, 1999; McBride, Corpe, & Wing, 1995; Mills, 1999; Segall & Moreno, 1999). 컴퓨터화 능력적응검사는 종전의 지필검사와 전혀 다른 개인 중심

의 검사 형태로서, 교수·학습에서의 개인교습과 같이 교육평가적 측면이나 윤리적 측면에서 매우 바람직한 평가방법으로 미래지향적 평가방법이라 할 수 있다. 개인 존중의 기본 철학에 기초하여 제시된 문항들 중에 학생이 선호하는 문항을 선택하여 응답하는 학생중심의 자기선택검사(self adaptive test)(Rocklin & O'Donnell, 1987)도 제안되었다.



〔그림 1〕 컴퓨터 능력적응검사의 알고리즘

컴퓨터화 능력적응검사나 자기선택검사는 개인중심의 검사로서 컴퓨터화 능력적응검사의 경우는 학생들의 사전능력을 분석하여 첫 문항부터 학생능력 수준에 부합하는 문항을 제시하고 정답 여부에 따라 쉬운 문항이나 어려운 문항이 제시되므로 학생들에게 적절한 긴장감과, 문제를 해결하려는 도전감, 문제를 풀고 난 뒤의 성취감, 이를 통한 긍정적 자아개념의 형성으로 개인의 정의적 행동특성 발달에도 바람직한 영향을 준다고 할 수 있다. Pitkin과 Vispoel(2001)은 자기선택검사가 능력을 정확하게 추정하기 위하여 컴퓨터화 능력적응검사보다 많은 문항이 필요하고 시간이 더 소요되지만, 검사 불안은 낮다고 보고하고 있다. 검사가 피험자에게 주는 검사 불안은 측정오차를 유발하므로 검사 불안을 최소화하는 것이 신뢰도를 높이는 방법이기도 하지만, 학생들의 정의적 행동특성 발달에도 긍정적 영향을 주기 때

문에 컴퓨터화 능력적응검사나 자기선택검사를 적용하는 것이 바람직하다.

3. 선다형 검사에서 수행평가로

개인의 능력을 평가하는 방법은 평가내용의 특성, 평가상황과 평가결과의 영향에 따라 달라진다. 인간에 대한 최초의 평가방법은 자연적 평가로서 구조화되어 있지 않았으며 있는 그대로를 평가하였으므로 인지적 능력의 평가는 대화법으로, 정의적 행동 특성은 자연관찰로 이루어졌으나 산업사회를 거치면서 교육평가의 행정적 기능이 강조되어 채점의 객관성을 보장하여야 하기 때문에 평가방법이 점차 구조화되었다. 구조화 정도에 따른 평가방법은 [그림 2]와 같다.

선택형 문항에 의한 평가 {구조화} — 수행평가 {비구조화}

진위형 | 선다형 | 배합형 | 괄호형 | 단답형 | 논술형 | 구술시험 | 수행평가 | 포트폴리오 | 참평가

인지	인지/정의/심동
앎	행함
학습결과	학습진행/결과
고정형 평가	개방형 평가
이분적 평가(하나의 정답)	다분적 평가(다양한 정답)
분석적 접근(analytical approach)	총체적 접근(holistic approach)
인위적 상황	실제적 상황
일회적 평가	지속적 평가
정적평가	동적 평가
행정적 기능	교수적 기능

〔그림 2〕 선택형 문항의 평가와 수행평가의 연속적 개념

평가방법의 구조화 경향은 서답형보다는 선택형 문항을 이용하게 하였고, 개인들은 주어진 답지에서 정답을 선택하는 수동적 자세를 갖게 되어 고등정신의 함양과 문제해결능력이 저하되는 결과를 가져오게 되었다.

지필검사에 의한 선다형 중심의 평가가 지니는 문제점을 해결하고자 1980년대 말 수행평가가 강조되었다. 수행평가란 습득한 지식, 기능이나 기술을 실제 생활이나 인위적 상황에서 얼마나 잘 수행하는지 혹은 수행할 것인지를 서술・관찰・면접 등의 방법을 통하여 수행과정과 결과를 종합적으로 판단하는 평가방법이다(최연희 외, 1999). 수행평가에 대하여 남명호(1995)는 '주어진 과제에 대하여 학생이 직접적이고 실질적인 수행의 평가'라 하였으며, 백순

근(1997)도 '학생 스스로가 자신의 기능이나 지식을 나타낼 수 있도록 산출물을 만들기 위하여, 행동으로 나타내거나, 답을 작성하도록 요구하는 평가방식'이라 정의하고 있다.

종전의 선택형 중심의 평가는 정답 여부만을 확인하는 평가방법으로 문제해결과정을 측정할 수 없을 뿐더러 개인들이 지닌 복합적 정신기능과 능력을 종합적으로 평가 할 수 없는 문제점을 지니고 있다. 이런 문제점을 극복하기 위하여 수행평가가 강조되었으며 이는 학생들이 지니고 있는 고유한 문제해결 방법과 고등정신능력을 발휘하게 하고 실제 생활에 적응하는 능력을 배양시키는 역할을 하고 있다. 나아가서는 인위적 상황에서의 수행평가보다는 실제 생활에서 실시하고 있을 내용을 평가하는 참평가(authentic assessment)를 실시함으로써 교육의 본질과 평가의 기능을 최대화 하려는 경향이다.

4. 검사에서 총평으로

산업사회로 발전하면서 검사는 고용·승진, 진학 등의 행정적 기능이 강화되고, 타당도와 신뢰도에 대한 요구가 강조되어 평가방법이 구조화되었다고 설명한 바 있다. 특히 기록으로 남겨야 할 필요성 때문에 지필검사가 실시되었고 그 결과로 개인을 평가하게 되었다. 그러나 개인의 특성을 검사 결과로만 평가하기에는 인간은 너무도 복합적인 개체이기에 인간을 평가하는 방법이 종합적인 평가체제인 총평으로 전환하고 있다. 총평(assessment)이란 여러 다양한 방법을 적용하여 종합적으로 평가하는 방법으로 사정이라고도 한다.

총평은 Murray(1938)가 사용한 용어로서 제2차세계대전 중 미 국방성에서 군사첩보요원을 선발하기 위하여 3일간 소집단을 구성하여 생활하게 하면서, 상황검사, 적성검사, 투사적 방법, 집중면담, 위기상황극복 평가 등을 실시한 예가 있다. 예전에는 중국에서 BC 1115년에 활쏘기, 말타기, 작문, 음악, 산술, 의식절차, 민원해결능력 등을 평가하여 관리를 선발하는 총평의 방법을 사용하였다.

현대사회가 요구하는 개인의 능력은 다양하고 보다 종합적인 특성들이어야 하므로, 개인에 대한 평가방법은 검사에서 탈피하여 종합적 평가체제인 총평의 시스템으로 변화되어야 할 것이다.

5. 일회적 평가에서 지속적 평가로

교육이 인간의 행동변화를 일으키는 행위라 할 때 인간의 행동변화는 한 순간으로 제한하지 않는다. 그러므로 교육평가는 종합적이고 지속적이어야 하는 특징을 지니고 있다. 교육평가에서 주의할 점 중 중요한 점은 개인에 대하여 표식을 다는 행위(labeling)이다. 한번 영재이면 영원한 영재라는 주장이 타당하지 않다는 것이다. 또한 개인의 어떤 능력이 열등하다

하여 평생 열등하지는 않다는 것이다.

개인의 노력이나, 교수·학습 환경에 의하여 개인들은 수도 없이 변화되고, 새로운 문제점에 직면하게 될 수 있으며, 다양한 변화를 가져올 수 있다. 거듭 변하는 것이 인간이고 생명체의 특성이기에 교육평가는 일회적 평가에서 끝나는 것이 아니라 지속적으로 평가되어야 한다. 평가방법을 체계화하고 평가의 주기를 정례화 하여 지속적으로 평가함으로써 개인의 변화정도와 경향 등에 대한 과학적 분석을 실시하는 것이 바람직하다.

개인뿐 아니라 교육과 관련된 모든 것, 즉 교육 프로그램이나 시설, 예산 등에 대해서도 지속적 평가를 실시할 때 교육 평가체제의 발전을 기대할 수 있다.

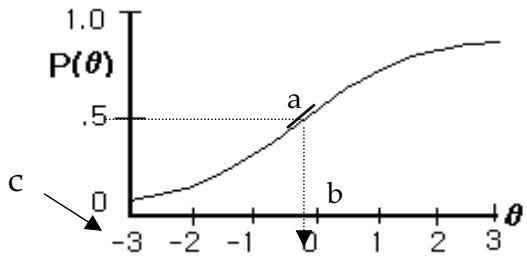
6. 고전검사이론에서 문항반응이론으로

인간의 능력을 측정함에 있어 측정의 결과를 계량화하려 함은 평가의 절차이기도 하다. 특히 상대비교평가의 경우, 측정한 점수가 필수적으로 요구되며 이를 통해 개인의 상대적 위치를 평가하게 된다.

인간이 지니고 있는 잠재적 특성을 분석하는 검사이론은 고전검사이론과 문항반응이론으로 나뉜다. 고전검사이론은 피험자가 얻은 관찰점수는 진점수와 오차점수로 합성되어 있다고 가정하며 진점수는 동일한 검사를 무한히 반복하여 얻은 점수의 평균점수로 계산한다. 문항의 난이도는 전체 응답자 중 정답자의 비율로 계산하고, 문항의 변별도는 문항의 정답여부와 총점의 상관계수로 추정한다. 문항의 배점은 문항의 형태가 같은 경우 동일한 점수를 부여하는 것이 일반적이나 문항의 중요도, 난이도 등에 따라 배점을 달리하기도 한다.

측정이론가들은 고전검사이론의 가장 중요한 가정인 진점수와 문항난이도 및 문항변별도에 대한 문제를 제기하였다. 고전검사이론은 같은 문항이라 할지라도 응답자 집단의 능력에 따라 문항난이도, 문항변별도 지수가 달리 추정되는 측정학적 모순을 지니고 있다. 개인의 능력 추정에 있어서도 쉬운 검사이면 능력이 과대 추정되고, 어려운 검사이면 과소 추정되며, 다른 문항일지라도 문항의 답을 맞힌 문항 수가 같으면 동점자로 분류되어 능력 추정의 정확성이 결여되고 있다. 각기 다른 문항들의 답을 맞혔다면 능력이 달리 추정되어야 하며, 선택한 오답지의 매력도에 따라서도 능력이 달리 추정되어야 한다.

고전검사이론이 가지는 측정학적 문제를 해결하기 위하여 문항특성의 불변성과 피험자 능력 추정의 불변성을 주요한 특징으로 하는 문항반응이론이 Lawley(1943)에 의하여 제안되었다. 문항반응이론에서는 개별 문항이 고유한 특성을 지니고 있다고 가정하며, 문항의 고유한 특성을 문항특성곡선으로 나타낸다. 문항특성곡선과 문항모수인 문항난이도와 문항변별도는 [그림 3]과 같다.



〔그림 3〕 전형적인 문항특성곡선

문항특성곡선은 피험자의 능력에 따라 문항의 답을 맞힐 확률을 말한다. 쉬운 문항이라면 문항특성곡선이 왼쪽에 위치할 것이고, 어려운 문항이면 오른쪽에 위치할 것이다. 문항반응이론에 의한 문항모수를 추정하는 수리적 방법은 Lawley(1943), Lord(1952), Birnbaum(1968), Baker(1992)등에 의하여 복잡한 계산 절차를 거쳐 발전되어 왔다. 문항반응이론에 의한 문항난이도는 문항이 답을 맞힐 확률이 .5이거나 (1+ci)/2에 해당하는 능력수준을 말하고 문항변별도는 문항난이도를 나타내는 문항특성곡선상의 점에서의 문항특성곡선의 기울기를 말한다.

문항반응이론에 의한 문항모수 추정이 수리적으로 간단하지 않으나 문항난이도와 문항변별도 등을 안정적으로 추정하여 준다는 장점이 있다. 즉, 능력이 낮은 피험자들이 응답한 자료를 가지고 문항을 분석하거나 능력이 높은 피험자들이 응답한 자료를 가지고 문항모수를 추정하더라도 두 추정치가 유사하게 추정됨으로써 문항특성의 불변성 개념을 유지할 수 있다. 뿐만 아니라 각기 다른 문항에 응답한 피험자들의 능력을 서로 다르게 추정함으로써 피험자들의 능력을 정확하게 추정하여 준다는 장점도 있다. 답을 맞힌 문항 수가 같다 하여 동점이 되는 것이 아니라 답을 맞힌 문항들이 다를 경우 해당 문항들의 모수에 근거하여 능력이 달리 정확하게 추정된다는 것이다. 또한 다른 응답지를 선택하였다면 오답지 선택에 따라 능력을 다르게 추정하는 정교함을 지니고 있다.

문항반응이론은 맞고, 틀리고의 이분적 분류가 되는 이분 문항반응이론(dichotomous IRT)에서, 여러 개의 답지나 응답을 분석하는 다분 문항반응이론(polytomous IRT)으로 발전되었을 뿐 아니라, 두 가지 이상의 특성을 측정하는 검사에도 적용 가능한 다차원 문항반응이론(multidimensional item response theory)으로까지 발전되고 있다. 또한 문항반응이론은 컴퓨터화 능력적응검사의 시행을 위한 필수적 이론이 되고 있다. 그 이유는 집단의 특성에 따라 영향받지 않는 문항모수가 문제은행에 저장되어 있어야 개인의 능력을 정확히 추정할 수 있기 때문이다.

Ⅲ. 개인화 종합평가시스템 구축과 적용

1. 개인화 종합평가시스템 구축

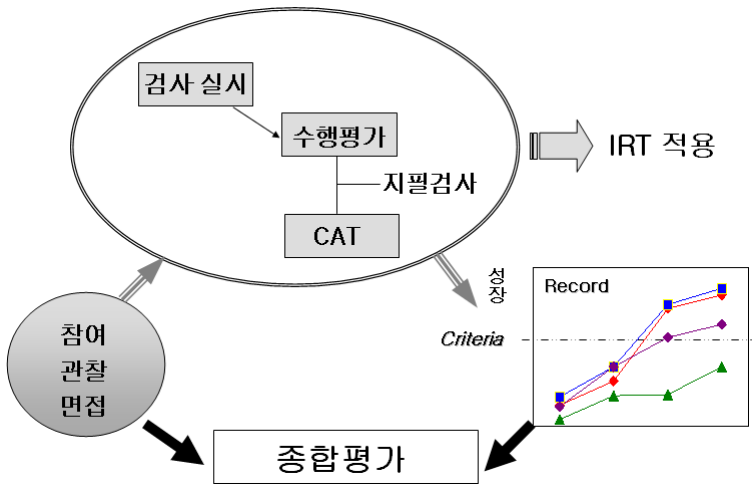
교육평가의 본질적 철학은 개인존중과 개인중심에 기인함으로써 평가체제가 성장참조평가로, 획일적 지필검사보다는 개인능력에 부합하는 컴퓨터화 능력적응검사로, 선택형 중심의 선다형 평가에서 수행평가로, 종합적 평가인 총평으로, 그리고 이를 위해서 일회적 평가가 아니라 지속적 평가로 변화되어야 한다고 주장하였다. 이와 같은 평가체제를 지원하기 위하여 고전검사이론보다는 과학적인 문항반응이론을 사용할 것을 제안한다.

이상에서 언급한 여섯 가지의 교육평가 방법 혹은 평가체제는 상호배타적(mutually exclusive)이거나 독립적인 것이 아니다. 문항반응이론의 적용 부분을 제외한 다섯 가지의 평가방법의 변화는 상호보완적이며 통합적으로 구안되어야 하며 그럴 때 교육평가의 기능을 극대화하는 시너지 효과를 얻을 수 있다.

인간의 특성을 평가하기 위한 방법으로 지필검사가 컴퓨터화 검사로 변화되는 것은 현대 문명의 흐름이다. 기능이나 기술의 습득 정도를 평가하기 위한 수행평가를 실시하고자 할 때에도 컴퓨터화 검사를 이용할 수 있다. 컴퓨터를 이용하여 인위적 혹은 가상적 상황을 제시하고 개인의 기술이나 기능의 정도를 측정하여 개인의 수행능력을 평가할 수 있다. 컴퓨터화 검사를 이용하면 평가문항이나 평가도구를 언제든지 사용할 수 있도록 준비할 수 있으므로 지속적인 평가가 가능하다는 장점도 있다. 또한 컴퓨터를 이용하면 개인에 대하여 수집한 자료를 종합적으로 관리하고 분석할 수 있는 체계를 갖출 수 있으므로 지속적으로 수집된 검사결과와 평가결과를 가지고 개인에 대한 총평이 가능하다. 이 자료에 근거하여 상대비교평가 뿐 아니라 절대평가, 나아가 성장참조평가도 실시할 수 있게 된다.

평가의 새로운 방향이 검사에서 총평으로 전환되고 있으므로 발전된 평가방법의 통합을 이루기 위하여 컴퓨터를 이용한 종합적인 평가체제 구축이 필요하고 이를 뒷받침하기 위하여 문항반응이론의 적용 뿐 아니라 최신 이론의 적용을 극대화하여야 할 것이다. 문항반응이론은 이분적 응답분석 모형의 수준을 넘어 다분적 응답분석, 나아가 다차원 문항반응이론도 적용되어야 할 것이다.

본 연구에서 개인화 종합평가시스템이라 함은 개인의 인지적, 정의적, 심동적 특성을 종합적으로 평가하기 위하여 지필이나 컴퓨터화 검사, 측정, 수행평가, 관찰 등의 다양한 방법을 이용하여 총평을 실시하고 개인의 성장·변화뿐 아니라 능력 정도를 종합적으로 평가하는 시스템이라 정의할 수 있다. 개인화 종합평가시스템은 [그림 4]와 같다.



〔그림 4〕 개인화 평가시스템 개요

예전의 지필검사 대신에 컴퓨터를 이용하여 검사를 실시하고, 선택형 문항의 경우 컴퓨터화 능력적응검사를 실시하고, 평가내용에 따라 수행평가를 활용한다. 검사나 평가결과를 컴퓨터에 저장하여 개개인에 대한 데이터베이스를 구축하면서 지속적으로 검사나 평가를 실시하게 되면 개인의 변화·성장을 분석할 수 있으며 자격증 부여를 위한 준거의 도달 여부 등을 종합적으로 평가할 수 있다. 개인의 특성을 다양한 측면에서 분석하기 위하여 실시한 면접이나 관찰결과를 검사결과와 종합하면 총평을 이용한 개인에 대한 종합적 평가가 가능하다. 이와 같이 컴퓨터를 기반으로 한 종합적 평가시스템이 구축되면 체계적이고 과학적이며 종합적으로 개인의 특성을 평가할 수 있을 것이다.

2. 개인화 종합평가시스템의 적용

개인에 대한 통합적 평가시스템은 교육뿐 아니라 산업계, 나아가 모든 분야에 널리 적용될 수 있다. 평가의 내용, 그리고 평가의 상황, 평가결과의 활용에 따라 평가방법이 다양하고 적용 범위가 제한적일 수 있으나 수행평가 결과, 성장참조평가 결과, 컴퓨터화 능력적응검사, 검사보다는 개인의 특성을 종합적으로 평가한 총평의 결과는 개인을 평가할 수 있는 중요한 자료가 될 수 있다.

예를 들어 의사에게 자격증을 부여하고자 할 때, 선다형 중심의 지필검사 결과보다는 컴퓨터에 나타난 환자 증상에 대한 동영상을 보면서 실제로 치료하거나 제출한 치료계획서가 더욱 타당한 평가결과이며, 의사로서 미래의 행위를 더 정확하게 예측할 수 있을 것이다. 의료기술에만 의존하던 평가를 넘어서 의사로서의 사명감과 가치관, 도덕성, 성격, 대인관계

등을 포함하는 총평의 결과가 있다면 이 평가자료 역시 많은 정보를 제공하여 우수한 의료인을 선발할 수 있을 것이다. 전문직에 종사할 개인을 평가할 때, 개인이 현재 지니고 있는 특성보다 오래 전부터 지니고 있는 개인의 지적 특성이나 정의적 행동 특성에 대한 성장과 변화도 고려한다면 개인에 대한 타당하고 신뢰할 수 있는 평가를 실시할 수 있을 것이다.

상대비교평가, 지필검사, 선다형 문항에 의하여 자격증을 부여하는 제도보다 개인화 종합평가시스템은 우수한 의료 인력을 확보할 수 있어 국민의 건강 증진에 이바지할 수 있는 장점이 있다. 더욱 중요한 것은 의료 전문인이 되기 위하여 학생들이 의학 전문 지식뿐 아니라 의사로서 지니고 있어야 할 특성, 즉 정의적 행동 특성에 대한 준비도 하게 한다는 것이다.

새로운 평가방법을 적용한 개인중심의 종합적 평가시스템은 최근에 연구되고 있는 의학전문인, 법학전문인 선발 뿐 아니라 교원평가, 공직자 다면평가 등 모든 분야에 적용이 가능하다. 현재까지 진행되어 온 학생선발, 분류, 배치, 자격증 부여 등의 모든 분야에 개인화 종합평가시스템이 적용된다면 관련 분야의 발전을 이룰 것이다.

3. 개인화 종합평가시스템이 주는 시사점

개인화 종합평가시스템이 우리나라 교육에 주는 시사점은 다음과 같다. 첫째, 상대비교평가를 벗어나야 한다. 평가의 목적상 상대비교평가를 사용할 필요가 있다 하더라도 가능한 이를 지양하여야 한다. 상대비교 결과가 주는 정보의 의미가 없을뿐더러 경쟁을 교육의 당연한 윤리로 받아들이기 때문에 학교현장에서 나타나는 문제점은 심각할 수 있다. 2008학년도 대학입학전형자료로 사용될 고교 내신 9등급제도에 대한 전면적 재검토가 요청된다. 대학수학능력시험도 영역이나 과목별로 점수 분포가 정규분포가 아니기 때문에 9등급제는 사용하는 것이 합리적이지 않다(양길석, 2006).

둘째, 학교현장에서 사용하고 있는 선다형 지필고사 중심에서 점진적으로 수행평가, 나아가서는 컴퓨터화 능력적응검사로서 계속적 평가와 지적 평가(intelligent measurement)를 실시하여야 한다. 지적 평가는 학생들이 가지고 있는 잘못된 인지구조를 변화시키고 잘못된 문제해결 능력을 교정하여 주기 때문이다.

셋째, 대학들이 신입생을 선발함에 있어 개인의 특성을 종합적으로 분석하여 전형에 반영할 수 있는 신입생 선발 방법에 대한 심도 있는 연구가 이루어져야 할 것이다. 특정 모집단위별로 전공에 따라 학생에 대한 종합적 평가결과를 반영하는 대학입학전형제도를 수립하여야 할 것이다.

넷째, 전문인들에게 자격을 부여하기 위하여 개인화 종합평가시스템에 의한 평가정보를 종합적으로 이용하여야 할 것이다. 검사결과나 특정 기술이나 기능에 의하여 자격을 부여하기보다는 다양한 특성들을 평가하여 자격을 부여한다면 보다 우수한 전문 인력을 확보할 수 있

을 것이다. 그러므로 국가고시에 자격증을 부여하는 기준 설정의 재검토가 필요하다.

다섯째, 평가결과가 개인에 미치는 영향이 크므로, 개인화 종합평가시스템에 대한 연구와 모형 개발 등은 물론 평가와 관련된 새로운 이론에 대한 이해와 적용을 위하여 지속적인 연수가 이루어져야 할 것이다. 이를 위한 연수의 내용으로는 평가의 개념 뿐 아니라 컴퓨터화 검사, 컴퓨터화 검사를 위한 문제은행의 개발, 검사의 동등화, 문항반응이론 등을 들 수 있다.

Ⅳ. 결론 및 논의

개인화 종합평가시스템은 제안 수준에 있으며 실용화를 위하여 많은 연구가 실행되어야 한다. 이 시스템을 제안함은 인간에 대한 평가는 과학적이고 체계적이며 지속적이고 종합적이어야 하며 교육적이어야 한다는 전제이다. 개인화 종합평가시스템을 활용하기 위하여, 교육평가전문가는 물론이고 교과내용전문가, 그리고 컴퓨터의 하드웨어와 소프트웨어 전문가의 공동연구 개발이 필요하다. 물론 부분적 개발을 통한 시스템 통합의 방법이 있으나 이보다는 전체적 구도에서 시스템을 개발하는 것이 바람직하다. 이를 위해서는 교육뿐 아니라 산업계, 의료계 등 전문 인력 양성기관 등에서 많은 관심과 연구지원이 이루어져야 할 것이다. 나아가서는 국가 인적자원개발 차원에서 개인화 종합평가시스템 개발에 대한 지원이 요구된다.

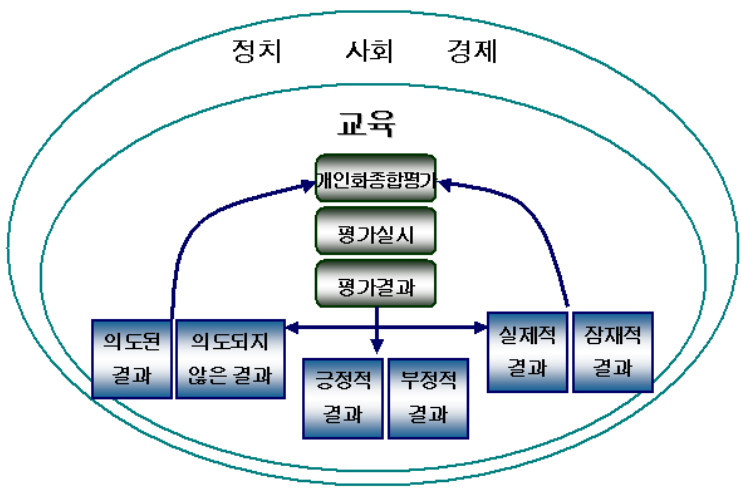
교육평가는 교육의 기능을 극대화 시켜야 한다. 그러나 그에 못지않게 가장 중요한 것은 윤리적 문제를 야기해서는 안 된다는 것이다. 컴퓨터화 검사든 수행평가든, 개인의 모든 자료를 수합하는 총평이든 평가의 대상인 개개인을 독특한 개체로서 인격적으로 다루어야 한다. 그러므로 교육평가는 인간을 유전적 입장보다는 환경론 입장에서 조망하여야 하며 학습자인 개인들은 무한한 잠재능력을 소유하고 있고 능력을 무한히 개발할 수 있다고 전제하여야 한다. 이런 전제하에 개인화 종합평가시스템이 구현되고 적용되어야 하며, 이 시스템이 개인을 평가할 때 긍정적이며 탈 권위적으로 교육활동의 보조적인 수단과 역할을 하여야 한다. 그러므로 개인화 종합평가시스템에 대한 학문적인 연구뿐 아니라 윤리적 연구도 병행하여야 한다.

개인화 종합평가시스템을 개발함에 있어 측정 내용도 간과할 수 없다. 인간을 평가함에 있어 평가내용과 평가방법은 개인 뿐 아니라 교수·학습의 내용과 방법 등을 구속하게 된다. 그러므로 교수·학습내용과 다른 내용들을 특이한 평가방법으로 평가하고 평가결과를 활용한다면, 교육내용은 특이한 평가방법과 평가내용에 의존하게 된다. 이를 측정선행교수(measurement-driven instruction: Messick, 1989)방법이라 한다. 평가가 교육내용이나 과정 등을

구속하는 측정선행교수 방법은 교육과정이나 교육내용을 인위적으로 변화시킬 수 있으나 고비용의 저 효과로 나타나는 경향이 높아 바람직한 교육평가 방법이라 할 수 없다. 그러므로 새로운 교육평가 방법이나 개인화 중심 평가시스템은 교수·학습을 도와주는 순기능이 강조되어야 하므로 교수선행측정 방법을 택하여야 한다.

인간을 평가함에 있어 더욱 중요한 것은 평가결과가 미치는 영향에 대하여 고려하지 않을 수 없다. 앞에서 제안한 평가방법의 전환이나 새로운 평가방법, 개인화 종합평가시스템을 적용하였을 때 [그림 5]와 같이 의도한 결과가 나왔는지, 의도하지 않은 결과 중 어떤 결과가 나왔는지, 긍정적이고 부정적인 결과가 무엇인지, 실제적 결과와 잠재적 결과가 무엇인지에 대한 검토가 추가될 때 평가의 윤리성은 보장된다. 물론 학교현장에 적용할 수 있는 가능성과 타당성도 검토되어야 할 것이다.

APA, AERA, NCME(1999)는 교육과 관련된 모든 현상을 평가할 때 결과타당도를 검토하여야 한다고 주장하고 있다.



〔그림 5〕 개인화 종합평가시스템의 결과 타당도

개인화 종합평가시스템 개발을 위하여 많은 이론적 연구와 실험적 연구가 이루어져야 한다고 본다. 성장참조평가 모형을 위한 이론정립 등과 더불어 개인맞춤평가, 그리고 컴퓨터화 검사에 대한 후속 연구가 필요하고, 특히 종합평가시스템 구축을 위한 각 분야별 구성요소와 요소 간의 연계, 그리고 데이터베이스 구축 및 관리방법에 대한 많은 후속 연구가 요구된다.

참 고 문 헌

남명호(1995). **수행평가의 타당성 연구**. 박사학위논문, 고려대학교.

백순근(1997). 수행평가의 이론적 기초. **수행평가의 이론과 실제**, 1-44. 한국교육평가학회 세미나.

성태제(1999). 교육평가 방법의 변화와 결과타당도에 대한 고려. **교육학연구, 37**(1), 197-218.

성태제(2001). **현대교육평가.** 서울: 학지사.

양길석(2006). 대학수학능력시험의 문제은행 구축 및 활용. **교육방법연구, 18**(2), 177-199.

유재봉(2007, 편집 중). **교육학개론**. 서울: 학지사.

정범모(1968). **교육과 교육학**. 서울: 배영사.

최연희 · 권오남 · 성태제(1999). **중학교 영어 · 수학 교과에서의 열린 교육을 위한 수행평가 적용 및 효과 분석**. 교육정책연구과제.

황정규(1998). **학교학습과 교육평가**. 서울: 교육과학사.

AERA/APA/NCME (1999). *Standard for Educational and Psychological* Testing. Washington DC: American Psychological Association.

Baker, F. B. (1992). *Item Response Theory: Parameter estimation techniques.* NY: Marcel Dekker Inc.

Bergstorm, B. A. & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Buchan (eds.), *Innovative in computerized assessment* (pp. 67-91). Mahwah, NJ: Lawrence Erlbaum Associates.

Binet, A. & Simon, T. H. (1905). Methods nouvelles pour le diagnostic du niveau intelletuel des anotmaux. *L'annee Psychologique, 11*.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (eds.), *Statistical theories of mental test scores. Reading*. MA: Addison-Wesley.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Processings of the Royal Society of Edinburgh, 18*, 1-11.

Lord, F. M. (1952). A theory of test scores. *Psychometrika Monograph, 7.*

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (ed.), *Computer-assisted instruction, testing, and guidance*. NY: Harper & Row.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement* (3rd ed.) (pp. 13-103). Washington DC: American Council on Education & National Council on Measurement in Education.

Mill, C. N. (1999). Development and introduction of a computer adaptive Graduate Records Examinations General Test. In F. Drasgow & J. B. Olson-Buchan (eds.), *Innovative in computerized assessment* (pp. 117-135). Mahwah NJ: Lawrence Erlbaum Associates.

Murray, H. A. (1938). *Explorations in personality*. NY: Oxford University press.

Oosterhof, A. (1994, 2001). *Classroom Application of Educational Measurement*. Merrill Prentice-Hall.

Pitkin, A. K. & Vispoel, W. P.(2001). Differences between self-adapted and computerized adaptive tests: A Meta-Analysis. *Journal of Educational Measurement, 38*, 235-247.

Rockin, T. R. & O'Donnell, A. M. (1987). Self adaptive testing: A performance improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79*, 315-319.

Segall, D. O. & Moreno, K. E. (1999). Development of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Buchan (eds.), *Innovative in computerized assessment* (pp. 67-91). Mahwah, NJ: Lawrence Erlbaum Associates.

• *논문 접수 : 2006년 10월 10일 / 수정본 접수 : 2006년 11월 15일 / 게재 승인 : 2006년 11월 27일*

ABSTRACT

A propose for Individualized Comprehensive Evaluation System

Tae-Je Seong
(Professor, Ewha Womans University)

A function of the educational evaluation makes human being grow individually. However, conventional standardized test, mechanical and group education or educational evaluations don't meet the goals of education and that these won't do good to individual growth and development or the growth of the society/nation.

There are changes for the educational evaluation from the norm-referenced evaluation to the growth-referenced evaluation, from a paper & pencil test to computerized adaptive test, from a multiple choice item to a performance assessment, from a test to an assessment, and from one-time measurement to continuous measurement in order to analyze the individuals' growth and changes.

This study suggested a comprehensive evaluation system focusing on the changes to the methods of educational evaluation and individual students according to the nature of educational evaluation. It's the new change from tests to assessment, which means a comprehensive evaluation system based on the computer is needed to integrated the advanced evaluation methods.

Tests are administered with the computer instead of the old paper-and-pencil resources. A selection type item, a supply type item or a performance assessment instrument is given as part of a computerized adaptive test. The performance assessment will given according to the evaluation contents. Once the test and evaluation results are stored in the computer and individual databases are set up to do ongoing tests and evaluations, it's possible to analyze individual changes and growth and check if they reached the criteria to obtain a certificate. Even assessment is possible when the test results are combined with the interview or observation results administered to analyze individual characteristics in more diverse aspects. Such a computer-based evaluation system will help evaluate individual characteristics in a more systematic, scientific, and comprehensive way. A comprehensive evaluation system can have a wide range of applications in education, industry, and other fields.

Since an individualized comprehensive evaluation system only is in the stage of

suggestion, there should be active research efforts among expertises in the field of educational evaluation, curriculum, and computer science, to put it to practical use.

Key Words : growth-referenced evaluation, computerized test, performance assessment, assessment, comprehensive evaluation