Communications of the Korean Statistical Society Vol. 15, No. 3, 2008, pp. 465–481

문항반응이론에서 피험자 능력 및 문항모수 추정 알고리즘 개발†

최항석[1], 차경준[2], 김성훈[3], 박정[4], 박영선[5]

요 약

문항반응이론(Item response theory: IRT)에서는 문항이 가지고 있는 특성을 기초로 피험자의 능력을 추정하고 동시에 각 문항별 문항특성곡선(Item characteristics curve: ICC)을 이용하여 문항모수를 추정하게 된다. 그러나 모수추정에 있어서 최대우도추정의 경우는 초기값과 다른 여러 문제들이 발생할 수 있다. 본 연구에서는 추정문제 해결방법의 대안으로 점근적 근사화 방법(Asymptotic approximation method: AAM)을 제안한다. 이는 자료의 수가 적거나 국소 변동이 있는 경우에 효과적인 추정방법이라고 할 수 있다. 이에 개발된 'Any Assess' 시스템을 모의실험을 통하여 신뢰성을 검정하였다.

주요용어: 문항반응이론(IRT); 문항특성곡선(ICC); 초기값; 점근적 근사화 방법(AAM).

1. 서론

문항반응이론(Item response theory: IRT)은 수학 및 통계모형으로서 교육 · 심리측정의 제 2 세대 이론으로 잘 알려져 있다 (Birnbaum, 1968; Baker, 1992). 일반적으로 IRT는 문항이 가지고 있는 특성을 기초로 피험자의 능력을 추정하며 각 문항별 문항특성곡선(item characteristics curve: ICC)은 실제로 검사정보를 추출하는데 유용하다.

Lord (1953)는 ICC의 문항 모수(item parameters)를 추정하기 위해 정규오자이브 모형(normal ogive model)에서 최대우도추정(maximum likelihood estimation: MLE)을 이용하였으며 로지스틱모형은 Birnbaum (1968)에 의해 연구 개발되었는데, 이후 MLE 과정을 위한 보다 단순화된 방법인 결합/조건/주변최대우도추정(joint/conditional/marginal maximum likelihood estimation) 등 베이지안 추정기법이 개발되어 피험자 수와 문항수가 적거나 잡음이 섞인 자료 등에서 신뢰성이 떨어지는 경우 또는 모수추정치의 왜곡현상(歪曲現象)을 줄이는데 획기적인 기여를 하였다 (Craven과 Wahba, 1977; Foutz, 1977).

---

† 본 연구내용은 2003년도 한국통계학회 춘계 및 추계학술대회에서 발표된 바 있음.

1) (135-710) 서울시 강남구 일원동 50번지 별관동 B121 삼성생명과학연구소 유전체연구센터, 연구원. E-mail: hangseok.choi@gmail.com
2) (133-791) 서울시 성동구 행당동 17번지 한양대학교 수학과, 교수.
3) (100-715) 서울시 중구 필동 3가 26번지 동국대학교 교육학과, 교수.
4) (609-735) 부산시 금정구 장정동 산 30번지 부산대학교 유아교육학과, 전임강사.
5) (133-791) 서울시 성동구 행당동 17번지 한양대학교 수학과, 강사. 교신저자: pppppys@hanyang.ac.kr

그러나 이러한 추정방법들은 내재된 여러 가지 문제점들을 가지고 있는데, 결합최대우도추정법의 문제점은 부분모수(incidental parameter)인 피험자 능력 모수가 구조적 모수(structural parameter)인 문항 모수에 영향을 준다는 것이다. 즉 결합최대우도추정법은 피험자수에 의하여 문항 모수 추정치가 변화되는 문제점을 지니고 있다.

또한, 대부분의 추정방법들은 3 모수를 추정하는 경우에 $3n + N$개의 모수 ($n =$문항수, $N =$피험자수)를 추정하여야 한다. 특히, MLE의 추정절차는 $3n + N$개의 미지 모수에 의하여 로그 우도 함수를 1차 편미분하여 0이 될 때 모수들의 추정치를 얻게 되며 Newton-Rapson의 반복적 절차를 통하여 문항 모수와 피험자 능력 모수를 추정하며 문항과 능력 모수 추정치의 증가분이 무시하여도 상관없을 정도의 값일 때 문항과 능력 모수의 추정은 종료된다. 그러나 이러한 과정에서 치명적인 문제는 '초기값(initial value)'에 있다. 우리는 이러한 문제를 해결하기 위한 대안으로 비선형 로지스틱모형을 단순선형회귀모형으로 근사화한 추정알고리즘을 제안하였다 (박영선 등, 2003a).

외국의 경우에는 이미 BILOG (Mislevy와 Bock, 1990)와 LOGIST (Wingersky 등, 1982) 등이 개발되어 상용화되어있으며, 최근에는 Lee와 Terry (2005)는 SAS Macros를 이용하여 피험자능력 및 문항모수를 추정하는 방법을 개발하였다.

본 연구에서는 저자 등이 개발한 'Any Assess'의 알고리즘을 2장과 3장에서 소개하고, 초기값과 점근적 근사화 모수추정(AAM) 알고리즘은 4장에서 다룬다. 그리고 프로그램실행과정은 5장에서, 6장에서는 모의실험자료를 통하여 BILOG와 비교함으로서 프로그램의 활용가능성을 검토하였다. 끝으로 7장에서는 토의와 결론이 제시된다.

2. 능력모수추정 알고리즘

피험자의 문항에 대한 반응은 각 문항을 맞추거나 틀리는 이분형(binary)이고 문항점수(item score) u_{ij}는 문항을 맞은 경우 '1', 틀린 경우 '0'으로 주어진다. 문항반응모형(ICC)은 피험자의 능력 함수로써 문항을 맞출 확률로 표현된다. 즉, 피험자 j의 능력값 θ_j에 대하여 문항 i를 맞을 확률과 틀릴 확률은 각각 아래와 같이 표현된다.

$$P\left(u_{ij}=1|\theta_j\right)=P_i\left(\theta_j\right), \qquad P\left(u_{ij}=0|\theta_j\right)=1-P_i\left(\theta_j\right).$$

그리고 피험자 능력추정기법으로 최대 우도(maximum likelihood: ML) 방법은 아래의 로그 우도 방정식을 최대로 하는 모수를 추정하는 것이다.

$$\log L\left(\theta_j|\underline{U_j}\right)=\sum_{i=1}^{n}\left\{u_{ij}\log P_i\left(\theta_j\right)+\left(1-u_{ij}\right)\log\left(1-P_i\left(\theta_j\right)\right)\right\}.$$

그러면 우도 방정식은 아래와 같고

$$\frac{\partial}{\partial\theta_j}\log L\left(\theta_j|\underline{U_j}\right)=\sum_{i=1}^{n}\frac{u_{ij}-P_i\left(\theta_j\right)}{P_i\left(\theta_j\right)\left\{1-P_i\left(\theta_j\right)\right\}}\cdot\frac{\partial P_i\left(\theta_j\right)}{\partial\theta_j}=0,$$

Fisher's scoring method에 의해 아래의 식이 수렴할 때까지 계산한다 (Baker, 1992).

$$\hat{\theta}_j^{(t+1)}=\hat{\theta}_j^{(t)}+\left\{I\left(\hat{\theta}_j^{(t)}\right)\right\}^{-1}\left\{\frac{\partial}{\partial\theta_j}\log L\left(\hat{\theta}_j^{(t)}|\underline{U_j}\right)\right\},$$

여기서 $I(\theta_j) = \sum_{i=1}^{n} a_i^2 P_i(\theta_j)\{1 - P_i(\theta_j)\}$이고 이 식을 Fisher의 정보함수라고 한다. 또한, MLE의 표준오차(standard error: SE)는 $\mathrm{SE}(\hat{\theta}_j) = \sqrt{1/I(\hat{\theta}_j)}$이다.

사후분포의 최대값(maximum a posteriori: MAP) 기법은 아래와 같이 베이즈 정리(Bayes' theorem)의 형태에 기초하여 추정한다.

$$g\left(\theta_j|\underline{U_j}, \underline{\xi_i}\right) \propto L\left(\underline{U_j}|\theta_j, \underline{\xi_i}\right) g(\theta_j),$$

여기서 $g(\theta)$는 사전분포(prior distribution)로서 평균과 분산이 각각 μ_θ, σ_θ^2인 정규분포를 따른다. 위의 식에 log를 취한 후 최대가 되는 능력모수를 MAP 추정치라 하고 그것은 아래의 식을 구한 해이다.

$$\frac{\partial}{\partial\theta_j} \log L\left(\underline{U_j}|\theta_j, \underline{\xi_i}\right) + \frac{\partial}{\partial\theta_j} \log g(\theta_j) = 0.$$

또한, MAP의 추정치 정확도 측도로서의 사후 표준오차(posterior standard error: PSE)는

$$\mathrm{PSE}\left(\hat{\theta}_j\right) = \sqrt{\frac{1}{J\left(\hat{\theta}_j\right)}}$$

이고 여기서 $J(\theta_j) = I(\theta_j) - \partial^2/(\partial\theta_j^2) \log g(\theta_j)$를 사후 정보(posterior information)라고 한다.

베이즈 EAP(expectation a posteriori) 추정 절차는 주변우도(MML) 추정방법에서와 같이 Gaussian 구적(quadrature)에 의해 근사적으로 계산한다 (Baker, 1992).

3. Any Assess 문항모수 추정 알고리즘

문항반응모형은 추정 대상인 문항 모수(item parameters)와 능력 모수(ability parameter)를 가진다. 임의의 문항(i)에 대해 그 모수는 변별도(power of discrimination: a_i), 난이도(item difficulty: b_i), 추측도(guessing: c_i)로 구성된다. 이들 모수들의 추정은 다음과 같은 단계로 이루어진다.

Step 1. 문항 반응 모수 a_i, b_i, c_i를 MML/EM으로 추정.

Step 2. 능력 모수 θ_j를 ML, MAP 또는 EAP로 추정.

위의 모수추정의 두 단계를 통합하여 결합최대우도(joint ML)라고 부른다. 실제로 모수추정은 3가지 가정 하에서 이루어진다. 첫째, 모든 피험자들은 서로 독립이고, 둘째, 모든 문항들은 서로 독립이며 셋째, 문항과 피험자는 서로 독립이다.

임의의 피험자(i)에 대해 능력 θ_j를 가진 반응점수의 패턴은 $\underline{U_j} = (u_{1j}, u_{2j}, \ldots, u_{nj})$으로 나타나고 이것의 확률은 아래와 같이 표현된다.

$$P\left(\underline{U_j}|\theta_j\right) = \prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} \{1 - P_i(\theta_j)\}^{1-u_{ij}}.$$

반응점수 패턴의 주변우도방정식은 $P(\underline{U_j}) = \int P(\underline{U_j}|\theta_j)g(\theta_j)\,d\theta_j$이다. 여기서, 능력 θ_j는 연속함수 $g(\theta_j)$의 분포를 따른다. 이 주변우도방정식은 Gaussian 구적 식

$$\overline{P}\left(\underline{U_j}\right) \approx \sum_{k=1}^{q} P\left(\underline{U_j}|X_k\right) A(X_k)$$

와 같이 근사적으로 표현되며, 여기서 X_k는 구적 점(quadrature point)이고 $A(X_k)$는 θ_j의 밀도함수 $g(X_k)$에 해당하는 가중치(weight)이다. 로그주변우도방정식은

$$\log L = \sum_{l=1}^{S} r_l \log \overline{P}\left(\underline{U_l}\right)$$

이다. 여기서 r_l은 N명의 피험자 중 같은 반응 점수패턴을 가진 빈도이고, S는 구분되는 패턴들의 그룹의 수이다. 위 식을 최대화하는 모수의 추정치를 주변우도추정치(marginal maximum likelihood estimate: MMLE)라 하고 그것은 아래의 로그우도방정식의 해이다.

$$\sum_{k=1}^{q}\left(\frac{\overline{r_{ik}} - \overline{N_k}P_i(X_k)}{P_i(X_k)(1-P_i(X_k))}\right)\frac{\partial P_i(X_k)}{\partial\begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix}} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

여기서 $\overline{r_{ik}} = \sum_{l=1}^{S}\{r_l u_{lj} P(\underline{U_l}|X_k)A(X_k)\}/\overline{P}(\underline{U_l})$이고 $\overline{N_k} = \sum_{l=1}^{S}\{r_l P(\underline{U_l}|X_k)A(X_k)\}/\overline{P}(\underline{U_l})$이다.

위의 식에서 문항 모수 a_i, b_i, c_i는 Fisher's scoring method와 EM algorithm (Bock과 Aitkin, 1981)을 이용하여 구하는데, 문항모수 초기값은 로지스틱모형을 단순선형회귀모형으로 근사화하여 추정한다 (박영선 등, 2003a). 또한, Expectation 과정에서 $\overline{r_{ik}}$, $\overline{N_k}$을 계산하고 Maximization 과정을 실행한 후 추정한다. Fisher's scoring method의 반복 계산 과정 중 문항모수의 분산이 0이 되는 경우에 추정치가 무한대로 증가하는 현상이 나타난다. 이러한 현상을 방지하기 위해 문항모수를 MMAP(marginal maximum a posteriori) 추정이라 부르는 베이즈(Bayes) 절차를 활용하여 추정한다. 베이즈 절차에 앞서 문항모수들에 대한 정보를 제공하는 사전분포로 문항의 기울기 a_i는 lognormal 분포, 하한 점근선 모수 c_i는 베타분포를 따른다 (Swaminathan과 Gifford, 1986). 여기서 문항모수는 $t_i = \log a_i$으로 변수변환이 된다.

문항모수 $\underline{\xi_i} = (a_i, b_i, c_i)$과 능력값이 따르는 분포가 가지는 모수 $\underline{\tau} = \left(\mu_\theta, \sigma_\theta^2\right)$의 사후분포(posterior)를 고려하면 아래의 베이즈 정리 형태이고 밀도함수 h_1, h_2는 각각 $\underline{\xi_i}$, $\underline{\tau}$를 따르는 모집단분포가 된다.

$$P\left(\underline{\xi_i}, \underline{\tau}|\underline{U_j}\right). \propto L\left(\underline{U_j}|\underline{\xi_i}, \underline{\tau}\right) h_1\left(\underline{\xi_i}\right) h_2\left(\underline{\tau}\right),$$

위의 식에 log를 취한 후 최대화하는 문항모수를 MMAP 추정치라 하고 그것은 아래의 식을 구한 해이다.

$$\frac{\partial}{\partial \underline{\xi_i}} \log L\left(\underline{U_j}|\underline{\xi_i}, \underline{\tau}\right) + \frac{\partial}{\partial \underline{\xi_i}} \log h_1\left(\underline{\xi_i}\right) = \underline{0},$$

$$\sum_{k=1}^{q}\left[\frac{\overline{r_{ik}}-\overline{N_k}P_i\left(X_k\right)}{P_i\left(X_k\right)\left\{1-P_i\left(X_k\right)\right\}}\right]\frac{\partial P_i\left(X_k\right)}{\partial\begin{pmatrix}t_i\\ b_i\\ c_i\end{pmatrix}}-\begin{pmatrix}\frac{t_i-\mu_{t_i}}{\sigma_{t_i}^2}\\ 0\\ \frac{\alpha-2}{c_i}-\frac{\beta-2}{1-c_i}\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix},$$

여기서 $\alpha=mp+1$, $\beta=m(1-p)+1$로 m은 사전정보를 따르는 가중치, p는 베타분포의 평균이다. 그리고 $\overline{r_{ik}}$, $\overline{N_k}$는 MML에서 제시한 식과 같고, EM을 이용하여 위의 식을 푼다.

4. 초기값과 점근적 근사화 추정알고리즘

모수추정에 있어서 첫 번째 문제는 MLE의 경우에 비선형 우도방정식을 풀어야만 하는데 이는 반복적인 Newton-Raphson기법을 적용하여 해결할 수 있다. 그러나 이 알고리즘은 초기값(initial value)에 민감하여 우도방정식의 해 근방의 초기값이 아니면 반복적인 과정에서 해에 수렴하지 못하는 단점을 가지고 있다.

또한, Samejima (1973)는 문항수가 적다면 하나이상의 근을 가질 수 있으며, 20문항이상이면 중다근은 발생하지 않는다고 하였다. 실제로 Foutz (1977)의 공리에 의하면 문항수가 충분히 많으면 근 θ의 유일성은 보장된다. 그러나 Looney와 Spray (1992)의 연구결과에 의하면, 표본이 크기나 검사길이가 늘어나면 측정의 표준오차는 감소하지만 연습이나 피로 그리고 검사를 반복함으로서 지역의 독립성가정이 지켜지지 않는 모순된 결과가 도출되기도 한다. 더욱이, 각 피험자가 비협조적인 경우와 피험자와 문항수가 적은 경우에는 더욱 모수추정에 신뢰할 만한 결과를 얻을 수 없다 (Mislevy와 Wu, 1988; Samejima, 1973).

요약하면 모수추정에 있어서 두 가지 문제는 초기값 문제와 문항수가 적고 피험자수가 작은 경우로 정리할 수 있다.

우리는 이러한 문제의 출발점을 ICC는 매끄러움(smoothness)을 갖는 단조증가함수이어야 하지만, 실제로는 그렇지 못한 이유가 측정오차로 인하여 국소 변동(local fluctuation)이 발생한다는 사실로부터 시작하였다. 이러한 추정문제의 특성은 관찰자료에서의 작은 변화가 근(모수)에서는 큰 변화로 나타난다고 할 것이다 (Craven과 Wahba, 1977; Stocking 등, 1973).

본 연구에서는 이러한 문항특성곡선의 내재된 문제점으로 인한 모수 추정문제에서 특히, MLE 추정절차에서의 'Newton-Raphson Algorithm'의 초기값(intial value)문제를 근사 해(approximate root)를 구하여 대체함으로써 여러 추정 난제들의 해결 대안으로 제시하였다 (박영선 등 2003a, 2003b).

일반적으로 모수 추정 시 초기값은 다음과 같이 설정되어 추정과정이 수행된다.

$$b_0=0.5-p\ (\text{해당문항의 정답률}),\qquad a_0=\frac{\operatorname{corr}\left(i,\ j\right)}{1-\operatorname{corr}\left(i,\ j\right)^2},$$

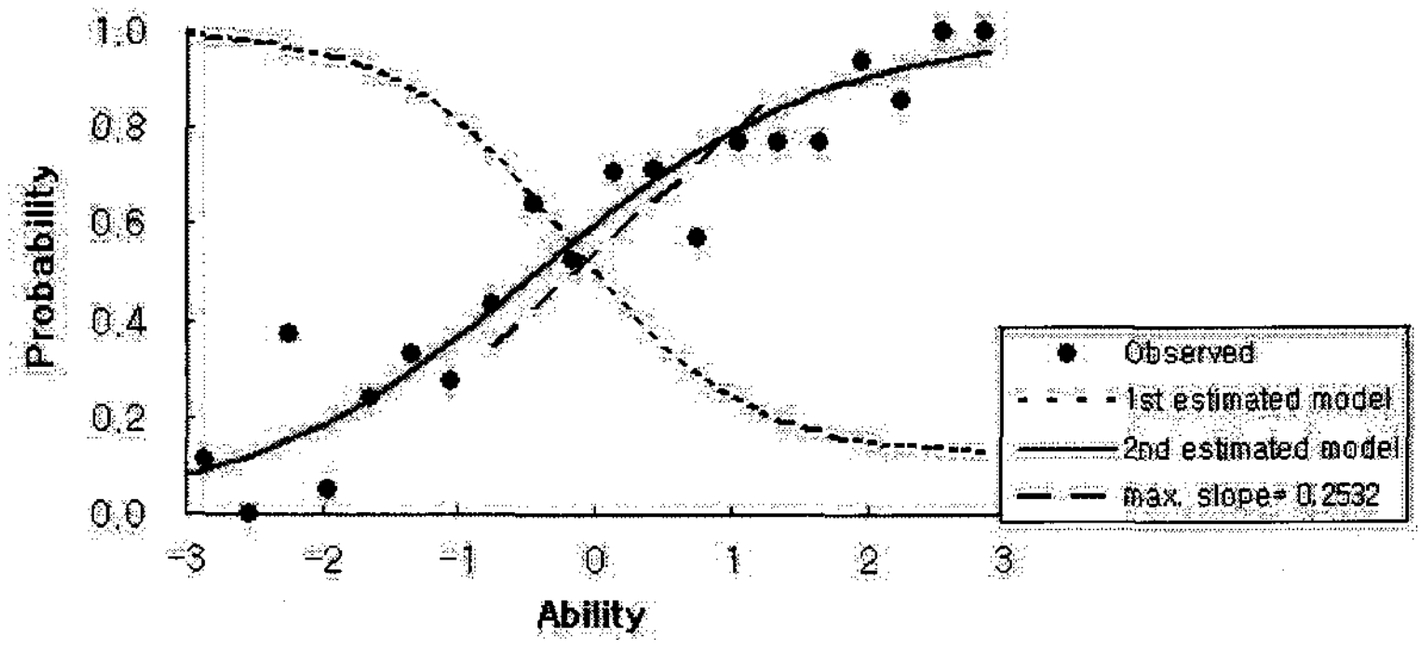

그림 4.1: 점근적 최대 경사의 추정을 이용한 ICC의 모수추정 예시 (1^{st} estimated model = 최적의 ICC 문항모수추정에 실패, 2^{nd} estimated model = 회귀기울기 (slope = 0.2532)에서 계산된 초기값 $(a_0, b_0, c_0) = (0.7109, 0.5810, 0.1620)$을 이용한 추정모형)

여기서 b_0는 고전검사이론의 문항난이도 p (해당문항의 정답률)에서 0.5를 기준으로 변형한 난이도의 초기값이며, 변별력모수의 초기값 a_0는 역시 고전검사이론에서 문항의 변별도로 이용하는 피험자(i)와 점수(j)의 상관관계 corr $(i,\ j)$를 이용한다. 또한, 추측모수의 초기값 c_0도 마찬가지로 고전검사이론에서의 추측도와 동일하게 설정된다 (이종성, 1990; 성태제, 1994).

본 시스템에서는 로지스틱 모형에서 단순선형 회귀 (독립 = 피험자능력, 종속변수 = 해당문항을 맞을 확률)를 이용한 점근적 최대 경사(β)의 추정과정을 거친 후 (박영선 등, 2003a), 먼저 하한 점근선(lower asymptote)인 초기 추측모수 c_0를 3개의 최소 능력값 $(\theta_0 < \theta_1 < \theta_2)$에 대해서 해당 문항을 맞출 확률 $P(\theta_i)$ (단, $i = 0,\ 1,\ 2$)의 산술평균값 $\{P(\theta_0) + P(\theta_1) + P(\theta_2)\}/3$으로 설정하고, 로지스틱모형의 변곡점이 $(b,\ (1+c)/2)$이고 최대경사 $\beta = 0.425 \times a \times (1-c)$이라는 사실로부터 변별력 모수의 초기값 a_0를 $\beta/\{0.425 \times (1-c_0)\}$로서 계산한다. 이때, 문항곤란도 b_0는 $(1+c_0)/2$이다. 즉, 3개의 모수 초기값

$$(a_0,\ b_0,\ c_0) = \left(\frac{\beta}{0.425 \times (1-c_0)},\ \frac{1+c_0}{2},\ \frac{P(\theta_0) + P(\theta_1) + P(\theta_2)}{3} \right)$$

을 이용하여 Newton-Raphson algorithm을 오차 < 0.1 범위에서 반복 시행한다.

그림 4.1은 상기한 점근적 최대 경사의 추정을 이용한 ICC의 모수추정 예로서, 초기 추정식 (1st estimated model)은 국소 변동(local fluctuation)으로 인하여 모수추정에 실패한 것을 보여주고 있다. 여기서 선형 회귀를 이용한 점근적 최대 경사(β) 즉, 회귀기울기(slope=0.2532)를 이용하면, 초기값을

$$\begin{aligned} (a_0,\ b_0,\ c_0) &= \left(\frac{0.2532}{0.425 \times (1-0.1620)},\ \frac{1+0.1620}{2},\ \frac{0.111+0.000+0.375}{3} \right) \\ &= (0.7109,\ 0.5810,\ 0.1620) \end{aligned}$$

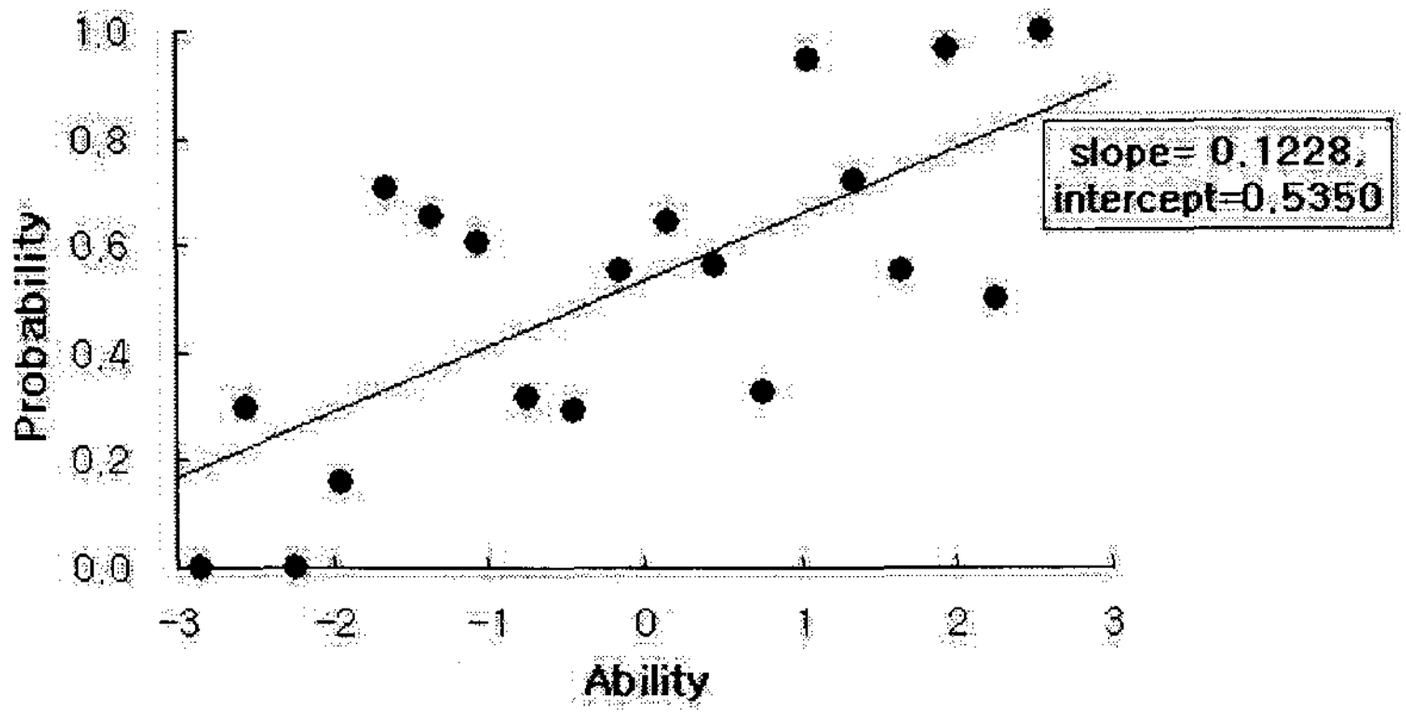

그림 4.2: 선형 회귀모형을 이용한 문항특성곡선의 근사 모수 추정 (실증사례: item = 13; $N = 1,000$ 중에서 $n = 200$개의 표본)

와 같이 계산할 수 있으며 이를 이용하면 ICC 추정모형 (2^{nd} estimated model)을 얻을 수 있다. 여기서 추정된 문항모수는 $(a,\ b,\ c) = (0.5536,\ 0.4253,\ 0.0001)$이고 최종적으로 추정된 ICC모형은 아래와 같이 표현 할 수 있다.

$$p(\theta) = 0.0001 + \frac{1 - 0.0001}{1 + \exp\left[-1.7 \times 0.5536 \times (\theta - 0.4253)\right]}.$$

또한, ICC 근사 모수 추정과정에서 초기값을 이용하였을 때에도 수렴하지 않는 경우 또는 ICC의 적합 자체가 의미가 없을 정도로 편의가 심한경우에는 직접 선형회귀모형을 적용하여 ICC 각 모수를 근사적으로 추정할 수 있다. 이는 피험자가 적은경우와 'bad items'의 경우에 문항정보 추출을 위한 알고리즘이다. 그림 4.2의 경우는 피험자 1,000명을 대상으로 실험자료 item=13에서 200명만을 추출하여 ICC 모수추정을 위 단계에서 초기값을 구하여 실시하였던 바, 수렴하지 않았다. 따라서 선형회귀모형이 $\hat{p}(\theta) = \hat{\alpha} + \hat{\beta}\theta = 0.5350 + 0.1228$로 추정되었다면 그림 4.2와 같이 모수 $a,\ b,\ c$를 다음과 같이 근사적으로 정의할 수 있다.

변별력모수는 $a = \beta/\{0.425 \times (1 - c)\} = 0.1228/\{0.425 \times (1 - 0.1666)\} = 0.3467$, 난이도모수는 $b = (1 + c)/2 = (1 + 0.1666)/2 = 0.5833$, 추측모수 $c = \hat{\alpha} + \hat{\beta} \times (-3) = 0.5350 + 0.1228 \times (-3) = 0.3467$로서 문항모수 $(a,\ b,\ c)$를 $(0.3467,\ 0.5833,\ 0.1666)$로서 근사적으로 추정할 수 있다.

상기한 점근적 최대 경사(β)의 추정과정의 실험결과와 자세한 절차는 박영선 등 (2003a)에서 참조 할 수 있다.

또한, 본 시스템에서의 점근적 근사화 모수추정(AAM) 알고리즘은 능력추정치로서 단지, 원 자료의 오답정보 (0 또는 1)를 이용한 z-점수를 이용함으로서 각 문항별로 알고리즘이 수행 될 수도 있다. 이는 소규모 검사에서 있을 수 있는 피험자 수와 문항수가 적은경우를 위한 실질적인 대안으로 활용을 기대할 수 있다.

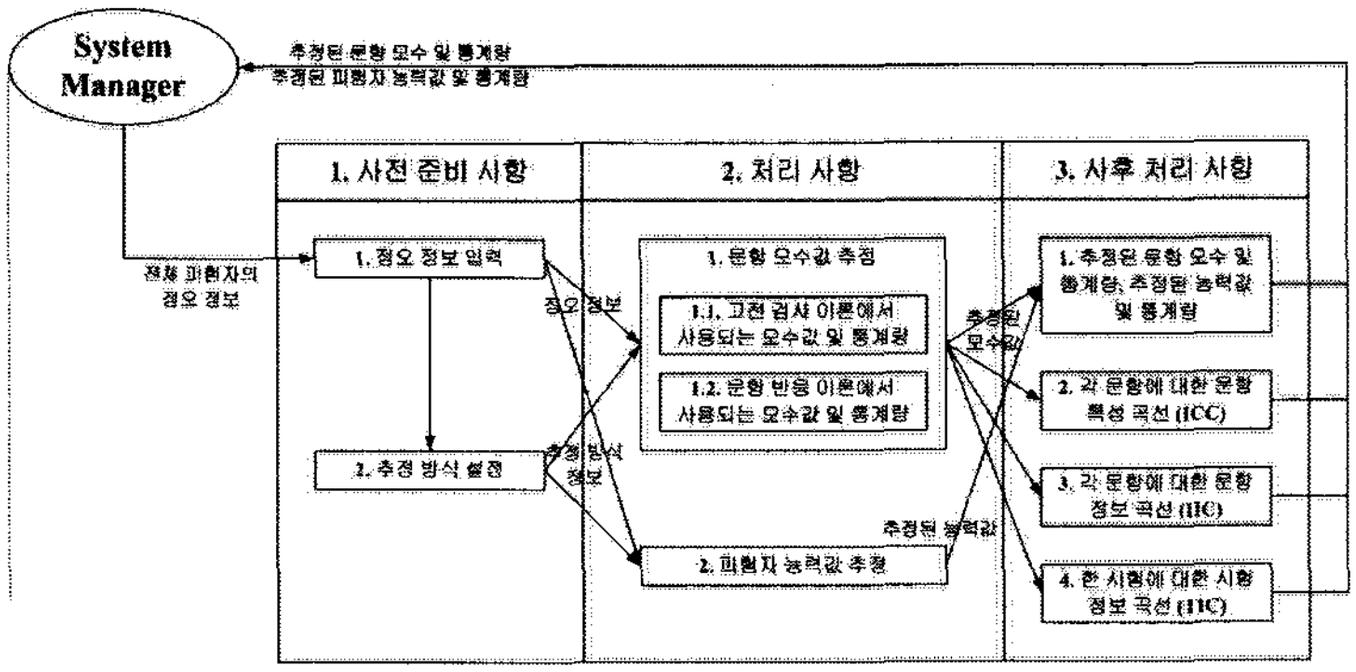

그림 5.1: Any Assess 시스템의 구조

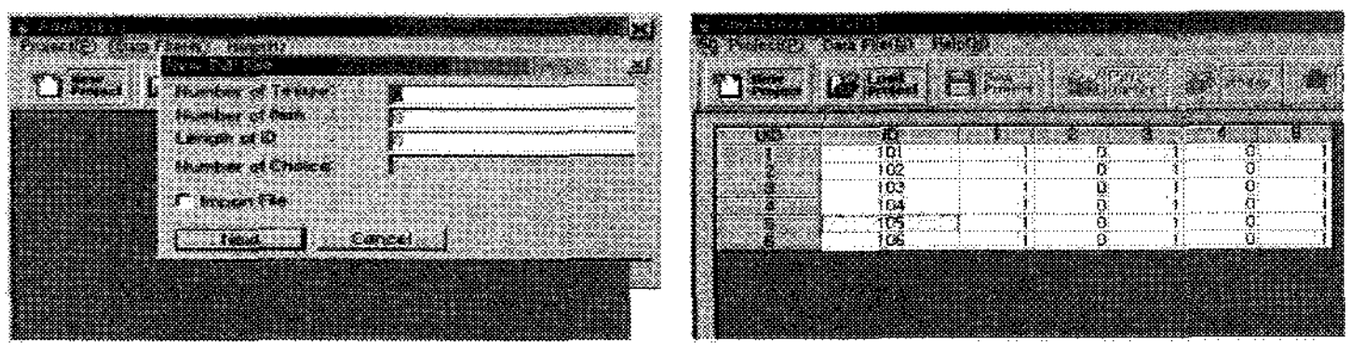

그림 5.2: Any Assess에서 자료 입력 (left)과 파일 불러오기 (right)

5. Any Assess 동작

'Any Assess' 전체 시스템의 구조는 그림 5.1과 같이 사전 준비사항, 처리 사항 그리고 사후 처리 사항인 3단계로 구성되어 있으며, 항시 피드백 과정 및 열람 등이 가능하도록 체계화 되어있다.

5.1. 자료구조

먼저, 분석을 하기 위한 대상 file이 필요하다. 'Any Assess'에서는 외부의 에디터를 이용하여 지정된 형식의 응답 파일을 만들거나 내부의 그리드 에디터(grid editor)를 이용하거나 혹은 Excel을 이용하여 대상 file을 만들 수 있다. 데이터의 형식은 크게 정오 형식과 피험자 응답 형식으로 나뉜다. 정오 형식 (*.tvf)은 피험자가 정답을 응답하였으면 '1'을 표기하고 오답을 응답하였으면 '0'을 표기하는 형식이며, 파일의 확장자는 'tvf'를 사용한다. 그림 5.2는 자료 입력과정으로서 Data File 메뉴에서 'New TVF'를 선택하면 새 TVF(정오 형식 데이터 파일)를 작성 할 수 있다 (left). 이후, 피험자 수(number of testee), 문항 수(number of item)와 식별자 길이(length of ID)를 입력하면, 입력 창이 열린다. 또한, 'New TVF'를 선택한 후, 'Import File'을 체크하면, 해당 Excel 파일을 불러올 수 있다 (right).

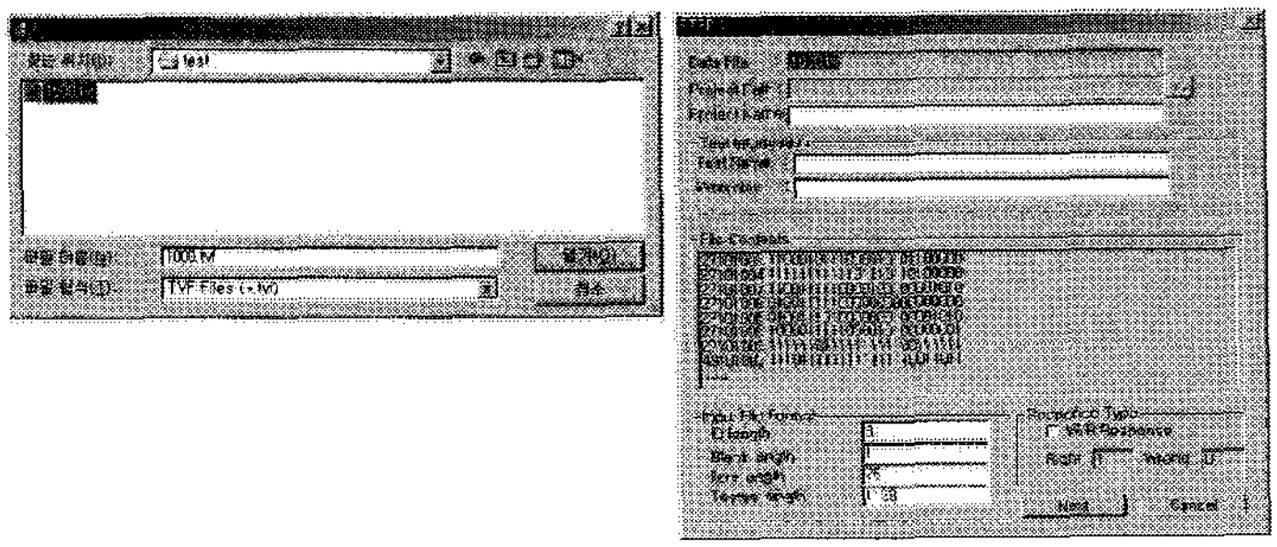

그림 5.3: Any Assess에서 파일의 지정 (left)과 내용 검색 (right)

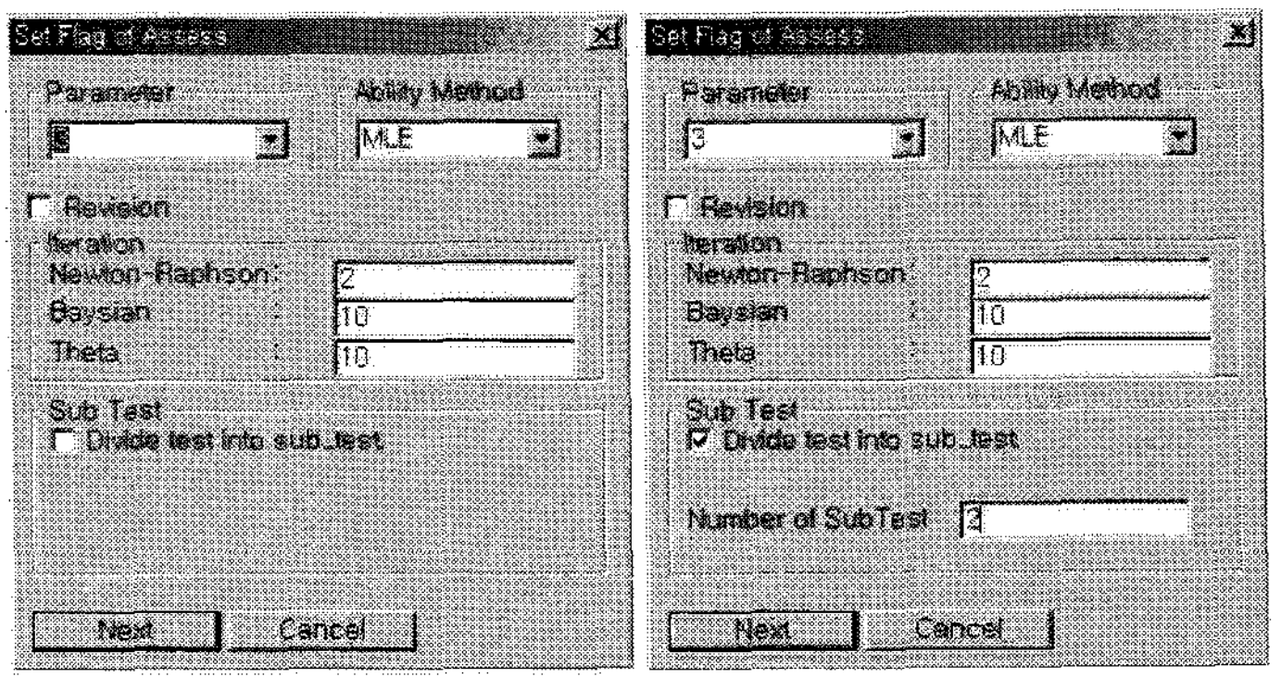

그림 5.4: Any Assess에서 분석방식과 하부 테스트 선택

5.2. 분석과정

분석에 관한 모든 정보는 프로젝트 파일에 의하여 관리된다. 프로젝트 파일은 따로 편집되지 않고 'Any Assess' 내부에서 관리되며 사용자는 내용을 반드시 알 필요가 없이 다이알로그를 선택하여 분석 방식을 지정하면 된다.

먼저, 그림 5.3과 같이 'New Project'를 선택하여 분석할 파일을 지정한다 (left). 파일을 지정하면 지정된 파일을 분석하여 식별자의 길이, 공백의 길이, 문항 수, 피험자 수 등을 알아낸다. 분석자는 프로젝트가 생성되기를 원하는 디렉토리와 프로젝트 명을 지정한다. 원한다면, 고사 명과 고사 시행자 명을 입력한다 (right).

다음으로, 그림 5.4에서와 같이 분석 방식을 지정한다. 모수 모형은 1-3모수 중 하나로 선택이 가능하고, 능력값의 추정 방식은 MLE, EAP 그리고 MAP 중 택일 할 수 있다. 'REVISION'을 선택하면 분석 후 보정 작업이 추가되고, 반복 횟수도 각각 지정할 수 있다. 시험이 여러 개의 하부 테스트로 이루어진 경우 'Sub Test'를 선택하여 모수 추정을 각기 수행할 수 있다. 특히, 모수추정 시 초기값은 4장에서 제안된 방법 (AAM)으로 계산되어 추정과정이 수행된다.

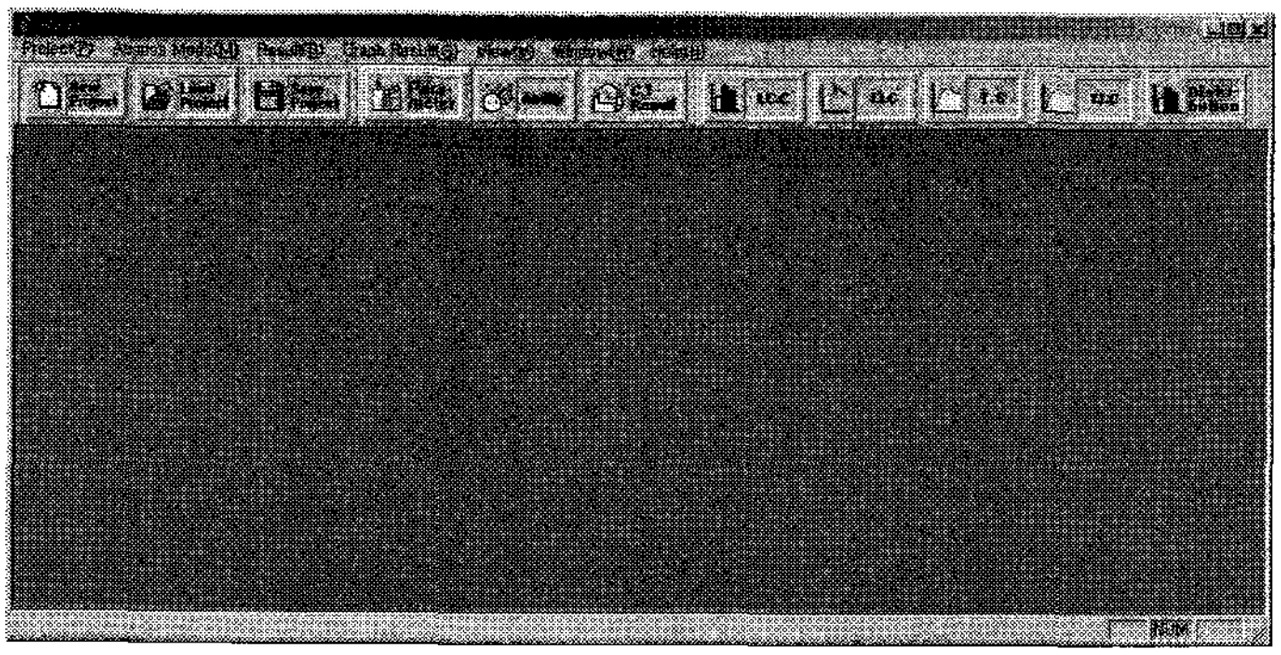

그림 5.5: Any Assess 분석 후의 프로그램 상태

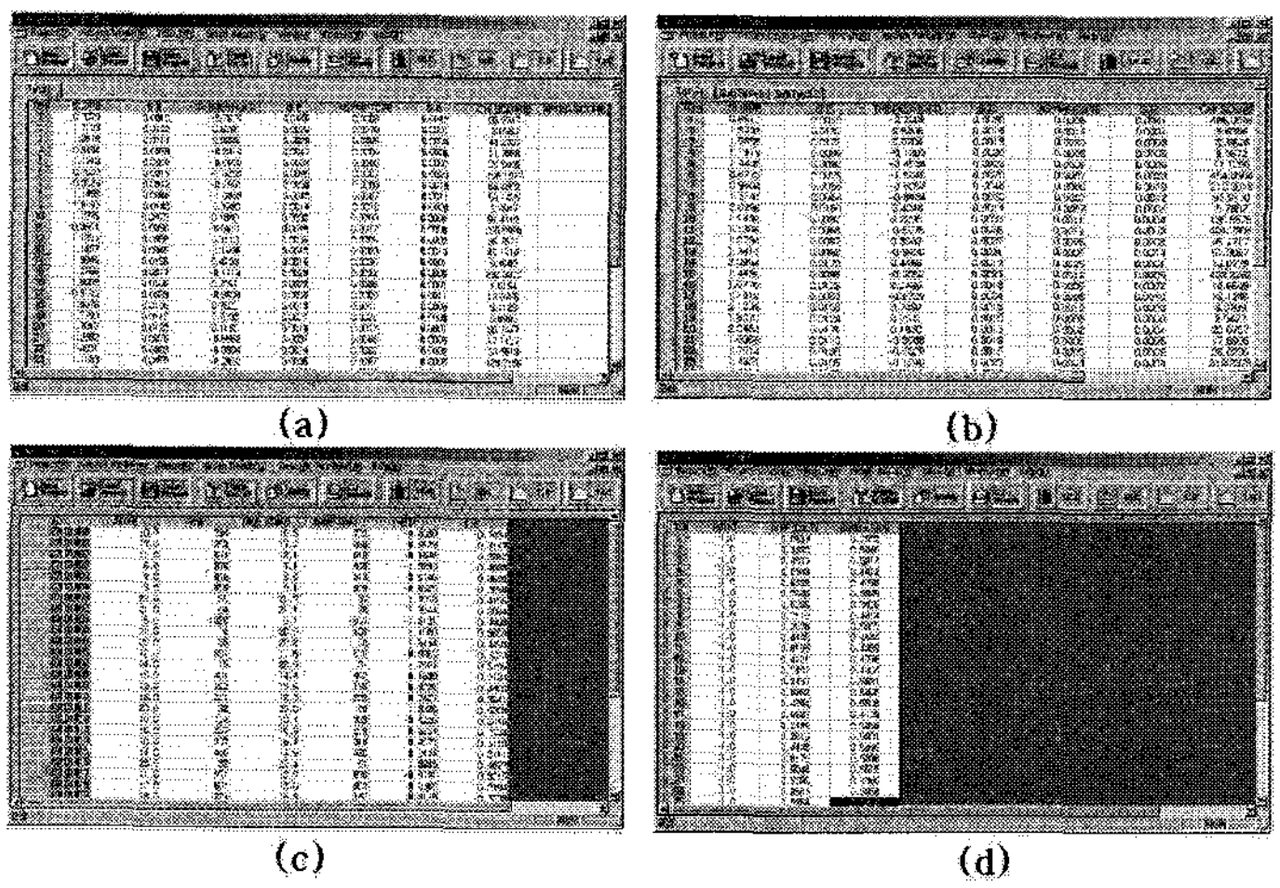

그림 5.6: Any Assess 분석결과 출력화면

5.3. 분석결과 보고 및 해석

'Any Assess'는 사용자의 편의를 생각하여 분석된 결과를 일반 텍스트가 아닌 그리드의 형태로 보기 쉽게 제공한다 (그림 5.5). 결과 출력과정으로 먼저 'Parameter'를 선택하면 문항반응이론에 의하여 분석된 문항의 모수 값이 그림 5.6과 같이 출력된다. 각 문항의 모수와 표준오차가 (a)와 같이 출력되고, 하부 테스트가 지정된 경우 각각의 하부테스트가 중첩되어 나타나고 가장 위에는 (b)와 같이 전체의 결과를 보여준다.

한편, 'Ability'를 선택하면 (c)에서와 같이 문항반응이론과 고전검사이론을 통틀어 피험자에 대한 분석결과가 나타나며, 피험자의 점수와 진점수 및 각각에 대한 순위, 그리고 능력 값이 출력 된다. 또한, 'C.T Result'를 선택하면 (d) 고전검사이론에 의한 문항의 난이도와 변별도가 계산된다.

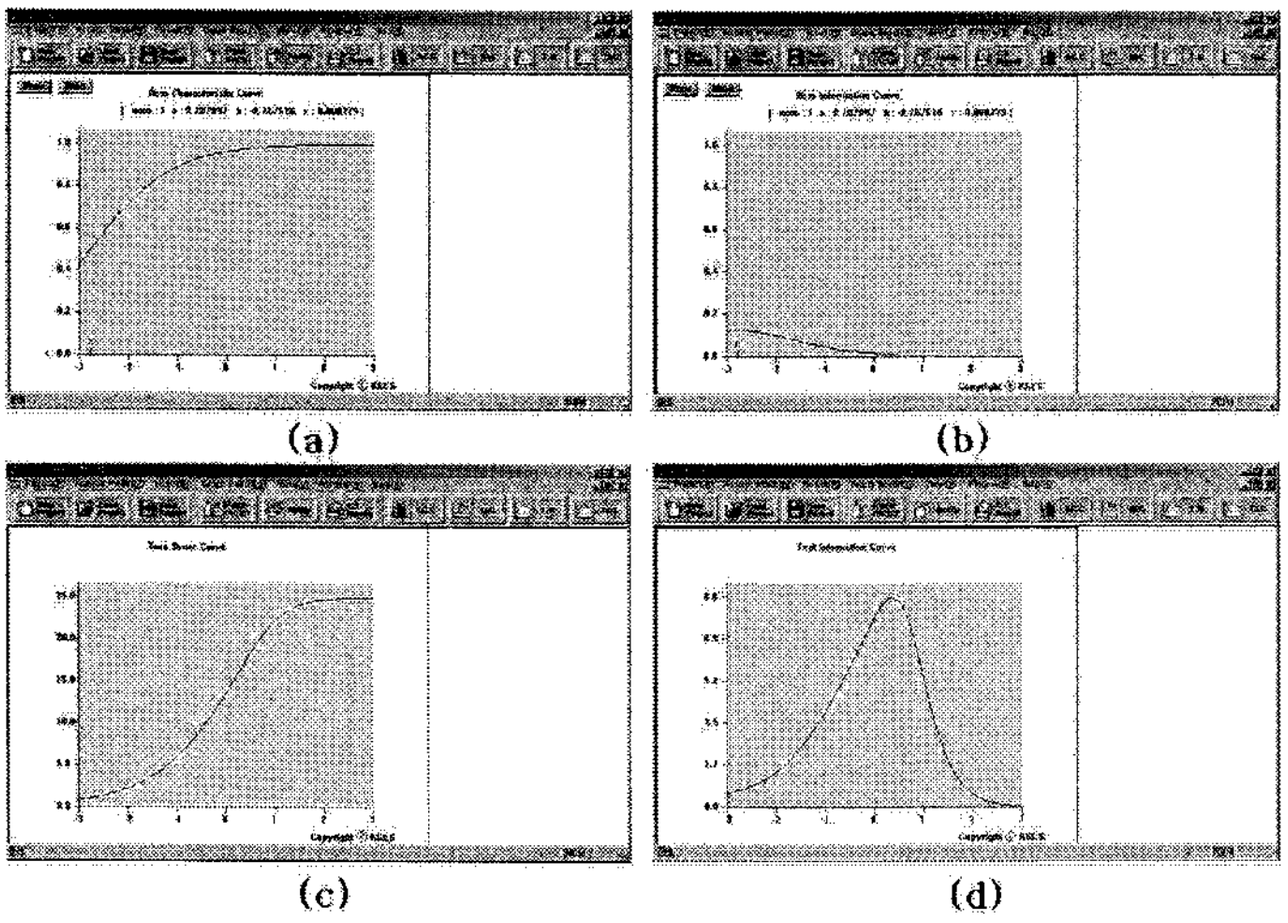

그림 5.7: Any Assess 분석결과 출력화면

그리고 분석 결과 그래프가 그림 5.7과 같이 출력되는데, (a)에서 'I.C.C.'를 선택하면 문항특성곡선(item characteristic curve)이, (b)에서 'I.I.C.'의 경우에는 문항정보함수(item information curve)를 그려준다. 또한, (c)에서 'T.S.'는 검사의 진점수 곡선(test score curve)을 분석해 주며, (d)에서 'T.I.C.'를 선택하면 검사정보함수(test information curve)가 나타난다.

6. 모의실험

'Any Assess' 검증을 위한 실험 자료는 RESGEN (Muraki, 2000) 프로그램을 이용 하여 생성하였는데, 문항수는 18문제, 피험자수는 1,000명으로서 실 문항모수와 피험자 능력을 각각 생성하여 외국에서 상용화된 BILOG (Mislevy와 Bock, 1990)와 더불어 비교하였다.

6.1. 문항모수 추정

1, 2-모수모형에서 실모수와 BILOG 그리고 'Any Assess'의 난이도와 변별력을 비교분석한 결과는 표 6.1과 같다. 그 결과를 살펴보면, 대부분의 추정모수는 실 모수와 비슷한 결과를 보였으며 BILOG의 추정결과와도 유사한 추정치를 보였다. 또한, MSE(mean squared error)를 구한결과, 1-모수모형에서 BILOG와 'Any Assess'는 각각 0.04600, 0.04692이었고, 2-모수모형에서는 먼저, 변별도에서 BILOG는 0.04558, 'Any Assess'는 0.03289이었으며 또한, 난이도에서는 각각 0.04641, 0.03973 로서 'Any Assess'가 비교적 낮게 나타났다. 3-모수모형에서도 전체적으로 추정모수는 실모수와 유사한 패턴을 보였으며 (표 6.2), MSE값의 비교에서도 BILOG와 큰 편차는 보이지 않았다.

표 6.1: 1, 2-모수에서 Real Parameter와 BILOG, Any Assess의 추정결과 비교

Item	1-모수			2-모수					
	Real	BILOG	Any Assess	Real		BILOG		Any Assess	
	b_i	b_i	b_i	a_i	b_i	a_i	b_i	a_i	b_i
1	−2.291	−2.227	−2.131	0.479	−2.291	0.484	−2.276	0.483	−2.319
2	0.027	0.051	0.065	0.421	0.027	0.386	−0.043	0.384	−0.094
3	1.125	1.209	1.193	0.529	1.125	0.611	1.102	0.599	1.059
4	−1.742	−1.597	−1.708	0.954	−1.742	0.951	−1.597	0.924	−1.708
5	0.430	0.376	0.280	1.077	0.430	1.020	0.376	0.689	0.280
6	1.353	1.302	1.146	1.014	1.353	0.941	1.302	0.992	1.146
7	−1.500	−1.597	−2.097	1.500	−1.500	1.267	−1.597	1.150	−2.097
8	0.098	0.215	0.189	1.588	0.098	2.127	0.215	1.816	0.189
9	1.673	2.048	1.855	1.542	1.673	1.035	2.048	1.339	1.855
10	2.291	2.142	2.801	0.479	2.291	0.467	2.142	0.418	2.642
11	0.027	0.291	−0.015	0.421	0.027	0.481	0.291	0.530	−0.015
12	1.125	1.835	0.991	0.529	1.125	0.648	1.835	0.433	0.991
13	1.542	1.431	1.364	1.154	1.542	1.150	1.431	1.156	1.364
14	0.430	0.524	0.425	1.077	0.430	0.926	0.524	1.002	0.425
15	1.353	1.425	1.168	1.014	1.353	1.270	1.425	0.696	1.168
16	−1.500	−1.551	−1.481	1.500	−1.500	1.210	−1.611	1.264	−1.602
17	0.098	0.131	0.142	1.588	0.098	1.660	0.068	1.706	−0.008
18	1.673	1.673	1.652	1.542	1.673	1.462	1.736	1.412	1.697

a_i = discriminating power; b_i = difficulty parameter.
Mean square error(MSE); 1-모수 (b_i); BILOG = 0.04600 *vs.* Any Assess = 0.04692,
2-모수(a_i, b_i); BILOG = (0.04558, 0.04641) *vs.* Any Assess = (0.03289, 0.03973)

6.2. 피험자 능력 추정

그림 6.1은 피험자 능력추정 알고리즘을 비교하기위해 정/오답정보가 (11011011001 1110110)인 경우에 실모수를 이용한 MLE, MAP, EAP 방법별 추정값 (MLE1, MAP1, EAP1)과 'Any Assess'에서 추정한 문항모수를 이용한 추정치 (MLE2, MAP2, EAP2)를 비교한 그림이다. 그 결과, 실문항모수를 이용한 추정치가 'Any Assess'를 이용한 것에 비해 실제 능력값 0.7879에 좀 더 근사하였으며 대체로 MLE의 경우보다 MAP, EAP 알고리즘이 좀 더 안정적인 경향이 있었다.

7. 결론 및 토의

현재 이용되고 있는 IRT의 개념은 잠재적 공간의 단일차원성(unidimensionality of latent space)과 지역독립성(local independence)을 가정하고, 부수적 모수(incidental parameter)인 피험자의 능력과 구조적 모수(structural parameter)인 문항모수의 관계를 확률함수로 규정하는 문항특성곡선(ICC)을 사용한다 (Andersen, 1970; Hambleton과 Cook, 1977).

표 6.2: 3-모수에서 Real Parameters와 BILOG, Any Assess의 추정결과 비교

Item	Real Parameters			BILOG			Any Assess		
	a_i	b_i	c_i	a_i	b_i	c_i	a_i	b_i	c_i
1	0.479	−2.291	0.221	0.453	−2.408	0.295	0.469	−2.871	0.168
2	0.421	0.027	0.197	0.414	0.122	0.323	0.487	0.149	0.073
3	0.529	1.125	0.273	0.918	1.604	0.414	0.273	1.153	0.091
4	0.954	−1.742	0.219	0.951	−1.597	0.294	0.924	−1.708	0.212
5	1.077	0.430	0.243	1.020	0.376	0.231	0.689	0.280	0.098
6	1.014	1.353	0.280	0.941	1.302	0.258	0.992	1.146	0.246
7	1.500	−1.500	0.050	1.267	−1.597	0.220	1.150	−2.097	0.063
8	1.588	0.098	0.187	2.127	0.215	0.240	1.816	0.189	0.195
9	1.542	1.673	0.278	1.035	2.048	0.273	1.339	1.855	0.283
10	0.479	2.291	0.221	0.467	2.142	0.311	0.418	3.242	0.070
11	0.421	0.027	0.197	0.481	0.291	0.279	0.530	−0.015	0.172
12	0.529	1.125	0.273	0.648	1.835	0.375	0.433	0.991	0.190
13	1.154	1.542	0.119	1.150	1.431	0.156	1.156	1.364	0.168
14	1.077	0.430	0.243	0.926	0.524	0.265	1.002	0.425	0.215
15	1.014	1.353	0.301	1.270	1.425	0.314	0.696	1.168	0.250
16	1.500	−1.500	0.050	1.600	−1.482	0.239	1.512	−1.568	0.069
17	1.588	0.098	0.187	1.276	0.219	0.220	1.544	−0.145	0.111
18	1.542	1.673	0.278	1.380	1.921	0.289	1.349	1.916	0.129

a_i = discriminating power; b_i = difficulty parameter; c_i = guessing parameter.
MSE; 3-모수(a_i, b_i, c_i);
BILOG = (0.05569, 0.06340, 0.00797) *vs.* Any Assess = (0.03381, 0.10718, 0.00768)

ICC를 이용한 모수추정에 있어서 대표적인 추정방법인 MLE는 비선형 우도방정식을 반복적인 Newton-Raphson기법을 적용하여 해결할 수 있으나 이는 초기값에 민감하여 우도방정식의 해 근방의 초기값이 아니면 반복적인 과정에서 해에 수렴하지 못하는 단점을 가지고 있다 (Kale, 1962). 더욱이, 각 피험자의 적절치 않은 반응 등 기타 원인을 알지 못하는 잡음(noise) 그리고 피험자와 문항수가 적은 경우에는 모수추정에 신뢰할 만한 결과를 얻기 힘든 실정이다 (Samejima, 1973; Cohen 등, 2001).

Bock과 Lieberman (1970)에 의해 제안된 주변최대우도추정(MMLE) 기법 역시, 표본수가 작은 경우 등에 한계를 가지고 있기 때문에 Wollack 등 (2002)은 마코프 연쇄몬테칼를로(MCMC) 방법 등을 대안으로 제시한 바 있다.

한편, 추정결과에 있어서는 IRT 자체 내 문제가 있을 수 있는데, Lord (1983)는 3모수 모형 하에서 피험자능력 모수를 안다고 가정했을 때 문항모수들이 어떻게 편의(bias)가 일어나는지를 살펴보았는데, 표준오차와 표본크기는 반비례하기 때문에 표집이 클 때는 편의 정도가 매우 작을 것이라고 결론지었다. Swaminathan과 Gifford (1986)는 3모수 모형 하에서 문항모수추정치의 일관성(consistency)을 조사하기 위하여 모의실험을 LOGIST (Wingersky 등, 1982)을 통하여 실험하였던바, 문항의 수와 표집 크기가 증가할수록 곤란도와 변별력 추정치는 실모수와 거의 유사한 값을 보인다고 하였다.

그리고 문항모수추정에서의 표준오차에 대한 연구는 Thissen과 Wainer (1982)가 대표

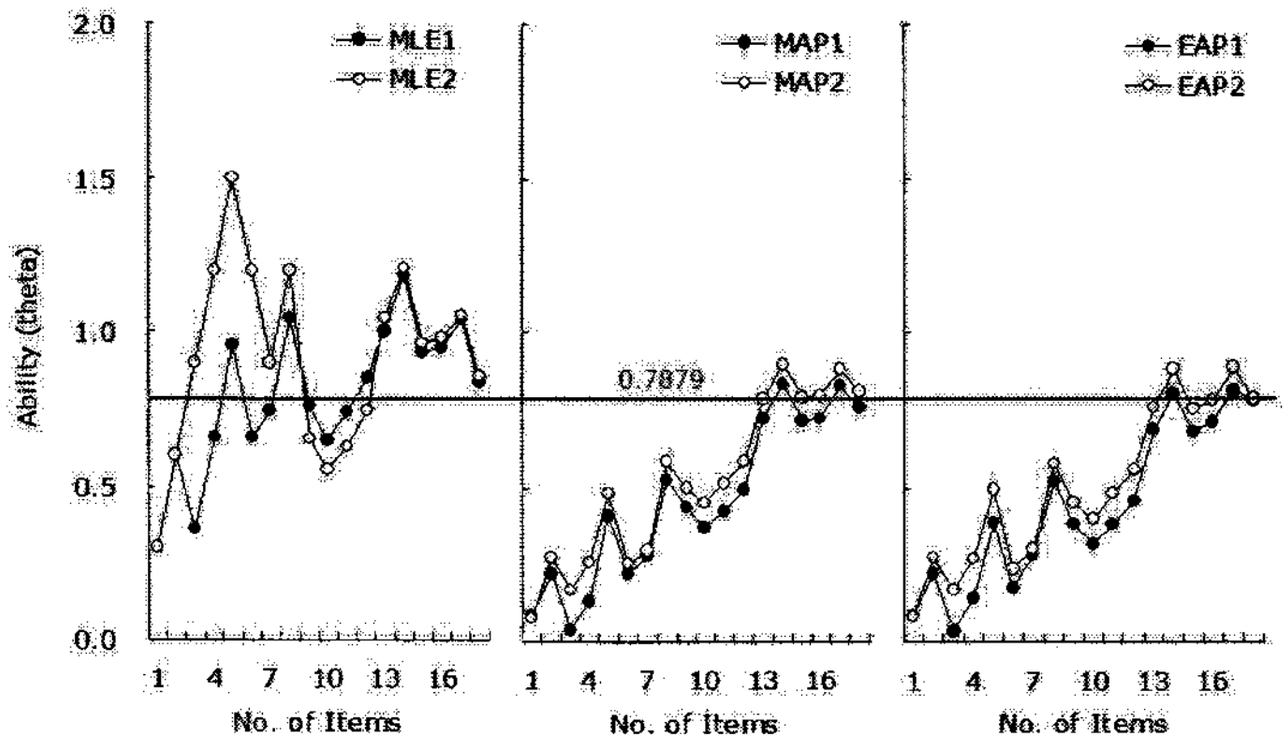

그림 6.1: 실모수(1)와 Any Assess(2)의 추정값을 이용한 MLE, MAP, EAP 능력추정 비교; Real ability = 0.7879. Mean squared error (MLE, MAP, EAP; N=1,000); BILOG=(0.05074, 0.04485, 0.01167) *vs.* Any Assess(1)=(0.04822, 0.04069, 0.00971) *vs.* Any Assess(2)=(0.05061, 0.04751, 0.01085)

적인데, 그들은 피험자의 능력모수가 알려져 있고 정규분포를 따른다는 가정 하에 1, 2, 3모수모형에서 문항모수의 점근적 표준오차를 조사하였다. 난이도의 표준오차는 3가지 모형 모두에서 오목하게 능력척도의 극단으로 갈수록 커지는 형태였으며 2모수모형이 3모수모형의 점근적 표준오차의 9%에 지나지 않았다. 이는 3모수를 적용하기 위해서는 대규모의 표집이 있어야 한다는 논지로서 3모수 모형에서의 모수추정의 어려움을 시사한다고 할 것이다.

IRT는 외국에서 뿐만 아니라 국내에서도 교육 및 심리측정 그리고 기타 여러 분야에서 활용과 그 가치를 인정받고 있다. 이에 본 연구에서는 첫째, IRT 문항모수 및 피험자능력 추정시스템 개발과 둘째, 일반적으로 존재하는 알고리즘 내의 초기값 문제 대안과 더불어 선형 근사모수 계산방법을 제안하였다. 이상에서 연구의 결론을 정리하면 다음과 같다.

첫째, 본 시스템은 외국의 상용화된 프로그램인 BILOG와 비교하여 전체적으로 큰 편차는 없는 것으로 나타났다. 피험자 능력추정방법에서는 MLE 방법보다는 베이지안 추정 알고리즘이 좀 더 실 모수에 적합한 추정치를 보였으며, 문항 모수추정 알고리즘에서는 1, 2-모수모형에서보다 3-모수모형에서 추정오차가 다소 높았으나 전체적으로 실 모수와 유사한 값을 얻을 수 있었다.

둘째, 본 시스템에서 새롭게 제안된 점근적 근사화 추정(AAM) 알고리즘은 개별 문항에 대한 모수추정을 수행할 수 있음을 보여주었다. 이는 ICC의 근사 모수추정치를 초기값으로 하여 다소 잡음이 포함된 경우와 피험자수가 적은 경우에 효과적이라고 할 수 있다.

국내 교육현장에서 가장 많이 활용하고 있는 외국의 상용화된 프로그램들 중에서 BILOG (Mislevy와 Bock, 1990)와 LOGIST (Wingersky 등, 1982)가 대표적인데, 이들은 다량의 문항과 많은 피험자를 대상으로 한 신뢰성 있는 모수추정에 역점을 두고 있다.

따라서 제안된 알고리즘인 비선형 ICC를 선형회귀모형으로 근사화 이론은 소규모 집단검사에 유용할 것으로 기대해 볼 수 있다. 물론 신뢰성 있는 모수추정은 될 수 없으나 현실적으로 대규모 집단검사가 불가능한 경우에 대안으로써 활용할 수 있겠다.

셋째, 단일문항평가가 이루어짐으로 해서 구조적으로 원활하고 단순한 'item-DB'구축이 가능할 수 있겠다. 또한, 신뢰성이 낮은 문항에 대해서는 'feedback'과정을 거쳐 신뢰도를 높일 수 있는, 재 추정방법을 사용한다면 검사의 직접적인 청정(purification) 효과도 가능하리라 기대된다.

따라서 향후에는 좀 더 정밀도를 높이는 문항별 잡음(noise) 제거, 자료의 보정 등의 부가적인 문제가 연구되어야 할 것이다.

감사의 글

본 논문을 심사하여 주신 익명의 심사자에게 감사를 드립니다. 특히, 한양대학교 통계정보분석센터의 연구원들과 (주)케이세스 임직원들에게도 감사의 말씀을 드립니다

참고문헌

박영선, 차경준, 장창원 (2003a). IRT 모수추정에서 초기값에 관한 연구, <한국통계학회 2003년 춘계학술발표회 논문집>, 7–12.

박영선, 진정언, 차경준, 이종성, 박정, 김성훈, 이원식, 이재화 (2003b). IRT에서 피험자 능력 및 문항모수 추정 알고리즘 개발, <한국통계학회 2003년 추계학술발표회 논문집>, 149–154.

성태제 (1994). 대학별고사를 위한 문항분석, 표준점수, 검사동등화, <한국통계학회 논문집>, **1**, 206–214.

이종성 (1990). <문항반응이론과 응용>, 대광문화사.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators, *Journal of the Royal Statistical Society, Ser. B*, **32**, 283–301.

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Technique*, Marcel Dekker, New York.

Birnbaum, A. (1968). Test scores, sufficient statistics, and the information structures of tests. In Lord, F. M. and Novick, M. R., *Statistical Theories of Mental Testscores*, Reading, Mass.: Addison & Wesley.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, *Psychometrika*, **46**, 443–459.

Bock, R. D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, **35**, 179–197.

Cohen, A. S., Bottge, B. A. and Wells, C. S. (2001). Using item response theory to assess effects of mathematics instruction in special populations, *Exceptional Childeren*, **68**, 23–44.

Craven, P. and Wahba, G. (1977). *Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation*. Technical Report No. 445, Madison, Department of Statistics, University of Wisconsin.

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations, *Journal of the American Statistical Association*, **72**, 147–148.

Hambleton, R. K. and Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data, *Journal of Education Measurement*, **14**, 75–96.

Kale, B. K. (1962). On the solution of likelihood equations by iteration processes. The multiparametric case, *Biometrika*, **49**, 479–486.

Lee, S. and Terry, R. (2005). IRT-FIT: SAS macros for fitting item response theory(IRT) models, *Presented at SUGI* 30^{th} *Conference in Philadelphia*.

Looney, M. A. and Spray, J. A. (1992). Effects of violating local independence on IRT parameter estimation for the Binomial Trials model, *Research Quarterly for Exercise and Sport*, **63**, 356–359.

Lord, F. M. (1953). Estimation of latent ability and item parameters when there are omitted responses, *Psychometrika*, **39**, 247–264.

Lord, F. M.(1983). Statistical bias in maximum likelihood estimators of item parameters, *Psychometrika*, **48**. 425–435.

Mislevy, R. J. and Bock, R. D. (1990). *BILOG 3: Item Analysis and Test Scoring with Binary Logistic Model*, Mooresville IN: Scientific Software, Inc.

Mislevy, R. J. and Wu, P. K. (1988). *Inferring Examinee Ability When Some Item Responses are Missing,* (Research Report 88-48-ONR), Princeton, N.J.: Educational Testing Service.

Muraki, E. (2000). *RESGEN: A Computer Program to Generate Item Response Vector*, Princeton: ETS.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory, *Psychomtrika*, **38**, 221–233.

Stocking, M., Wingersky, M. S., Lees, D. M., Lennon, V. and Lord, F. M. (1973). *A Program for Estimating the Relative Efficiency of Tests at Various Ability Levels, for Equating True Scores and for Predicting Bivariate Distributions of Observed Socres*, Research Memorandum 73–24, Princeton, N.J.: Educational Testing Service.

Swaminathan, H. and Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model, *Psychometrika*, **51**, 589–601.

Thissen, D. and Wainer, H. (1982). Some standard errors in item response theory, *Psychometrika*, **47**, 397–412.

Wingersky, M. S., Barton, M. A. and Lord, F. M.(1982). *LOGIST user's guide*, Princeton, NJ: Educational Testing Service.

Wollack, J. A., Bolt, D. M., Cohen, A. S., Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparision of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation, *Applied Psychological Measurement*, **26**, 339–351.

[2008년 2월 접수, 2008년 4월 채택]

Development of Estimation Algorithm of Latent Ability and Item Parameters in IRT†

Hangseok Choi[1], Kyungjoon Cha[2], Sunghoon Kim[3], Chung Park[4], Youngsun Park[5]

Abstract

Item response theory(IRT) estimates latent ability of a subject based on the property of item and item parameters using item characteristics curve(ICC) of each item case. The initial value and another problems occurs when we try to estimate item parameters of IRT(*e.g.* the maximum likelihood estimate). Thus, we propose the asymptotic approximation method(AAM) to solve the above mentioned problems. We notice that the proposed method can be thought as an alternative to estimate item parameters when we have small size of data or need to estimate items with local fluctuations. We developed 'Any Assess' and tested reliability of the system result by simulating a practical use possibility.

Keywords: Item response theory(IRT); item characteristics curve(ICC), initial value problem, asymptotic approximation method(AAM).

---

† This research was presented to members by us at 2003 The Korean Statistical Society Seminar.

1) Researcher, Center for Genome Research, Samsung Biomedical Research Institute, B121, Annex, 50, Irwon-dong, Gangnam-gu, Seoul 135-710, Korea. E-mail: hangseok.choi@gmail.com

2) Professor, Department of Mathematics, Hanyang University, 17 Haengdang-dong, Seongdong-Gu, Seoul 133-791, Korea.

3) Professor, Department of Education, Dongguk University, 26 Pil-dong 3-ga, Jung-gu, Seoul 100-715, Korea.

4) Full Time Lecturer, Department of Early Childhood Education, Pusan National University, San 30 Jangjeon-dong, Geumjeong-gu, Busan 609-735, Korea.

5) Lecturer, Department of Mathematics, Hanyang University, 17 Haengdang-dong, Seongdong-Gu, Seoul 133-791, Korea. Correspondence: pppppys@hanyang.ac.kr