아시아교육연구 12권 1호
Asian Journal of Education
2011, Vol. 12, No. 1, pp. 119-140.

문항모수 추정치에 기반한 문항 및 검사 신뢰도의 추정: 혼합유형문항검사 자료를 중심으로

강태훈(姜兌勳)*

논문 요약

이 연구에서는 다양한 교육 및 심리 검사의 결과에 대해서 문항반응이론(IRT: item response theory) 모형을 활용하여 문항 및 검사 신뢰도를 추정하는 방법을 소개한다. Dimitrov(2003)는 이분 문항으로 이루어진 검사 결과를 가지고 IRT 모수 추정치의 함수로서 문항 및 검사 신뢰도를 추정하는 방법을 제안하였는데, 이 연구에서는 이를 확장하여 다분 및 혼합유형(mixed format) 문항 검사에 모두 적용할 수 있는 일반화된 방법을 유도함으로써 보다 폭넓은 적용이 가능하도록 하였다. 이와 같은 방법론의 실제적 적용을 예시하기 위하여 각각 영어와 수학의 성취도를 측정하는 두 검사의 자료가 사용되었다. 신뢰도 추정치들을 살펴본 결과, 문항 수준의 신뢰도는 각 검사의 제작 및 문항 평가에 있어서 유용한 정보를 제공해 줄 수 있을 것으로 나타났다. 두 실제 검사 자료 모두에서 IRT에 기반을 둔 검사 수준의 신뢰도는 Cronbach's α 계수에 비해서 약간 큰 값을 가지는 것으로 추정되었는데, 이러한 현상의 이유를 설명하기 위하여 두 신뢰도 추정 방법의 관계 및 차이가 심층적으로 논의되었다.

주요어: 문항반응이론, 혼합유형 문항 검사, 문항 신뢰도, 검사 신뢰도

I. 서론

문항반응이론(IRT)은 검사 문항들에 대한 피험자의 수행을 묘사하기 위하여 고안된 여러 가지 수학적 모형들로 이루어져 있으며, 이를 통하여 문항과 피험자의 상호작용에 대한 효과적인 이해가 가능하다. IRT의 적용을 통하여 검사 결과에 관련하여 얻을 수 있는 정보는 매우 다양한

* 성신여자대학교 교육학과 조교수

데, 그 중에서도 문항과 피험자의 특성에 대한 추정이 가장 중요할 것이다. 그 외에도 주어진 능력 수준(θ 또는 trait)에서의 오차 변량이라든가 개개인의 진점수에 대한 추정, 그리고 문항 및 검사 수준에서의 신뢰도 추정 등 폭넓은 활용이 가능하다. 그럼에도 불구하고 검사 개발자나 제작자가 교육 및 심리 검사의 신뢰도를 구하는 경우, 자료 분석에 사용되는 측정 이론이나 통계적 모형에 관계없이 관습적으로 고전검사이론(CTT: classical test theory)에 바탕한 검사-재검사 신뢰도나 내적 일관성 신뢰도 등이 주로 쓰이는 경향이 있다.

IRT에 기반한 신뢰도를 사용하게 되면 CTT에서는 계산 불가능한 문항 수준의 신뢰도를 구할 수 있을 뿐만 아니라, 개별 피험자가 응답하는 문항들과 그 전체 숫자가 개개인마다 다를 수 있는 CAT(computerized adaptive testing) 하에서도 검사의 신뢰도를 효율적으로 계산해낼 수 있는 장점이 있다. 또한 측정학적 필요 혹은 실질적인 교육 및 심리 이론에 따라서 같은 구인(construct)를 재는 여러 개의 문항 집단 혹은 검사집(test booklet)이 필요할 때, 실제 자료를 이용하지 않고도 문항 모수와 피험자 능력 분포에 대한 정보만 있으면 각 검사집의 신뢰도를 추정하는 것이 가능하게 된다(Dimitrov, 2003). 이러한 특성은 하나의 문항 은행으로부터 여러 개의 검사를 만들어 낼 필요가 있는 상황 하에서 매우 유용하게 쓰일 수 있다. 물론 실제 검사 실시 없이 이렇게 모수 추정치만으로 예측된 문항 및 검사 신뢰도가 얼마나 정확한 것인가는 사용된 문항반응모형의 적합도나 IRT의 기본가정이 제대로 충족되고 있는가에 많은 영향을 받을 것이다. 하지만 검사정보함수나 능력모수 추정의 오차라는 유용한 정보에 더하여 검사 도구가 얼마나 일관되고 정확하게 피험자 특성을 측정하는 지에 대한 단일한 지수를 실제 검사 실시 없이도 구할 수 있다면, 이는 시간과 비용의 절약뿐만 아니라 교육 현장의 실무자나 연구자 사용하게 될 측정 도구 혹은 척도의 특성을 보다 효율적으로 파악하는 데에 도움이 될 것이다. 또한 때로는 검사정보함수보다는 신뢰도의 개념이 측정 및 검사이론에 대한 소양이 부족한 다른 연구자나 일반인과 의사소통하는 데에 있어서 보다 유용할 수 있다.

아래에서는 우선 Dimitrov(2003)가 제안한 방법을 바탕으로, 이분 문항 검사 결과로부터 추정된 문항 모수를 가지고 문항 및 검사의 신뢰도를 계산하는 방법을 소개하며, 이어서 이를 다분 문항 검사 결과로 일반화하는 방법이 제시된다. 혼합유형(mixed format) 검사의 신뢰도는 이들 두 방법을 각 문항의 유형에 맞게 통합함으로써 추정이 가능하다. 다음으로는, 이와 같은 IRT를 이용한 신뢰도 추정 방법을 두 가지 실제 자료에 적용하여 실질적인 유용성을 확인하고자 하였고, 덧붙여서 문항 신뢰도를 보다 의미 있게 해석하기 위한 방법을 제안하였다. 마지막으로, Cronbach's α 신뢰도와의 관계 및 차이를 논의하였다.

Ⅱ. 문항 및 검사 신뢰도의 추정 방법

이 논문에서는, 이분 문항 자료를 위해서 Birnbaum(1968)의 3모수 로지스틱 모형(3PLM: 3 parameter logistic model)을 사용하였고 다분 문항 자료를 위해서 Muraki(1992)의 일반화 부분점수 모형(GPCM: generalized partial credit model)을 사용하였다. 또한 혼합 유형 검사 분석을 위해서는 이 두 가지 모형이 함께 고려되었다.

1. 이분 문항 검사의 경우

3PLM 하에서 θ_j를 가지는 피험자 j의 문항 i에 대한 정답 확률은 식 (1)과 같다.

$$P_i(\theta_j) = \gamma_i + (1-\gamma_i)\frac{\exp[\alpha_i(\theta_j-\beta_i)]}{1+\exp[a_i(\theta_j-\beta_i)]} \quad (1)$$

여기서 α_i, β_i, 그리고 γ_i는 각각 문항 변별도, 곤란도, 그리고 추측도를 나타내는 모수이다. 이때, θ의 확률밀도분포를 $\phi(\theta)$로 보면 기대되는 문항 i의 오차 변량과 진점수 변량은 각각 다음 식 (2)와 (3)과 같이 문항 및 피험자 모수의 함수로 표현된다.

$$\sigma^2(e_i) = \int_{-\infty}^{\infty} P_i(\theta)[1-P_i(\theta)]\phi(\theta)\,d\theta \quad (2)$$

$$\sigma^2(\tau_i) = \int_{-\infty}^{\infty} P_i(\theta)^2\phi(\theta)\;d\theta \;-\left[\int_{-\infty}^{\infty} P_i(\theta)\phi(\theta)\;d\theta\right]^2 \quad (3)$$

오차변량을 식 (2)와 같이 구할 수 있는 이유는, 일차원성 가정을 갖는 IRT를 적용하는 상황 속에서 주어진 θ에서의 정답확률 $P_i(\theta)$는 베르누이 분포의 모수와 같이 기능하므로 그 무선적 변량을 $P_i(\theta)[1-P_i(\theta)]$로 구하는 것이 가능하기 때문이다. 여기에 능력분포를 가중치로 삼아 적분함으로써 한 문항에서의 오차변량을 구한 것이다. 이를 바탕으로 n개의 문항을 가진 전체 검사의 오차 변량과 진점수 변량을 구하려면 각각 다음 식 (4)와 (5)와 같이 계산할 수 있다(Lord & Novick, 1968; May & Nicewander, 1994).

$$\sigma^2{}_e = \sum_{i=1}^{n} \sigma^2(e_i) \tag{4}$$

$$\sigma^2{}_\tau = \int_{-\infty}^{\infty} [n\overline{P}(\theta)]^2 \phi(\theta)\ d\theta \quad - \left[\int_{-\infty}^{\infty} n\overline{P}(\theta)\phi(\theta)\ d\theta \right]^2 \tag{5}$$

여기서 $\overline{P}(\theta)$는 주어진 θ에서 계산된 n개의 (i=1,2,...,n) $P_i(\theta)$ 값들의 평균인데, 다르게 말해서 이 평균값에 n을 곱한 $n\overline{P}(\theta)$은 주어진 θ값에서의 진점수 추정치 즉 검사특성곡선(TCC: test characteristic curve)과 같다고 볼 수 있다. 위의 식 (2), (3), (4), 그리고 (5)를 바탕으로 문항 및 검사 신뢰도는 각각 다음 식 (6)과 (7)을 통해 구해질 수 있다.

$$\rho_{ii} = \frac{\sigma^2(\tau_i)}{\sigma^2(\tau_i) + \sigma^2(e_i)} \tag{6}$$

$$\rho_{xx} = \frac{\sigma^2{}_\tau}{\sigma^2{}_\tau + \sigma^2{}_e} \tag{7}$$

CTT 하에서 '정확성 계수(coefficient of precision)'로서의 신뢰도는 개념적인 공식일 뿐 실제로는 검사-재검사신뢰도 혹은 동형검사신뢰도와 같이 두 검사 결과간의 적률상관계수나 반분신뢰도 혹은 Cronbach's α와 같은 내적일관성 지수로 추정된다. IRT에서는 앞의 식 (4), (5) 등에서 알 수 있듯이 오차점수 변량과 진점수 변량을 문항반응모형을 이용하여 직접적으로 추정하기 때문에 CTT에서 불가능한 정확도 계수의 추정이 가능하게 된다.

2. 다분 문항 검사의 경우

한 문항 i에 두 개 이상의 반응 범주(x=0,1,...,mi)가 있을 때 이를 다분 문항이라고 하는데, GPCM 하에서는 이러한 각 범주 x에 대한 피험자 j의 반응 확률을 다음 식 (8)과 같이 모형화할 수 있다. 이러한 값들을 문항범주특성곡선(ICCC: item category characteristic curves)이라고 부르는데, 한 문항 당 범주의 수가 mi + 1개이므로 역시 mi + 1개의 ICCC가 존재하게 된다. mi는 문항 i의 총 범주 수에서 1을 감한 값이 되며, 이를 통해 문항마다 범주의 수가 다른 경우를 표현할 수 있다.

$$P_x(\theta_j) = \frac{\exp\sum_{k=0}^{x}\alpha_i[\theta_j - (\beta_i - \tau_{ik})]}{\sum_{y=0}^{m_i}\exp\sum_{k=0}^{y}\alpha_i[\theta_j - (\beta_i - \tau_{ik})]} \quad (8)$$

여기서 α_i와 β_i는 각각 문항 i의 변별도와 곤란도를 나타내며, τ 는 각 범주의 위치 모수를 의미한다. 식 (8)의 모형을 제대로 식별(identification)하여 문항 및 피험자 모수를 추정하기 위해서는 $\tau_{i0} = 0$, $\sum_{k=1}^{m}\tau_{ik} = 0$ 그리고 $\exp\sum_{k=0}^{0}\alpha_i[\theta_j - (\beta_i - \tau_{ik})] = 1$과 같은 세 가지 조건이 필요하다.

다분문항 자료의 경우에도 역시 θ의 확률밀도분포를 $\phi(\theta)$로 보면, 기대되는 문항 i의 오차변량은 다음 식 (9)와 같이 문항 및 피험자 모수의 함수로 표현된다. 이는 식(2)의 일반화된 형태로 볼 수 있으며 m=1일 때 즉 가능한 반응이 0과 1 뿐일 때에는 식 (2)와 같게 된다.

$$\sigma^2(e_i) = \int_{-\infty}^{\infty}\left[\sum_{k=0}^{m_i}k^2P_k(\theta) - \left(\sum_{k=0}^{m_i}kP_k(\theta)\right)^2\right]\phi(\theta)\,d\theta \quad (9)$$

이분문항 자료를 위한 식 (1)에 대응되는, 다분IRT(PIRT: polytomous item response theory) 하에서의 문항 특성 곡선(ICC: item characteristic curve)을 반응기대값곡선(REVC: response expected value curve)이라고 부를 수 있는데(Roberts et al., 2000), 이는 다음 식 (10)과 같이 구할 수 있다. 또한, 이를 이용하여 식 (11)과 같이 PIRT 하에서 기대되는 문항 진점수 변량을 구할 수 있다.

$$R_i(\theta) = \sum_{k=0}^{m_i}k\,P_k(\theta) \quad (10)$$

$$\sigma^2(\tau_i) = \int_{-\infty}^{\infty}R_i(\theta)^2\phi(\theta)\;d\theta \quad -\left[\int_{-\infty}^{\infty}R_i(\theta)\phi(\theta)\;d\theta\right]^2 \quad (11)$$

이를 바탕으로 n개의 다분 문항을 가진 전체 검사의 오차 변량과 진점수 변량을 구하려면 각각 다음 식 (12)와 (13)과 같이 계산할 수 있으며 이는 이분 문항 검사 결과를 위한 식 (4)와 (5)에 대응하는 개념이다.

$$\sigma^2{}_e = \sum_{i=1}^{n} \sigma^2(e_i) \tag{12}$$

$$\sigma^2{}_\tau = \int_{-\infty}^{\infty} \left\{ \sum_{i=1}^{n} R_i(\theta) \right\}^2 \phi(\theta)\ d\theta \quad - \left[\int_{-\infty}^{\infty} \left\{ \sum_{i=1}^{n} R_i(\theta) \right\} \phi(\theta)\ d\theta \right]^2 \tag{13}$$

여기서 $\sum_{i=1}^{n} R_i(\theta)$는 주어진 θ에서 계산된 검사 진점수의 추정치라고 볼 수 있다. 식 (10), (11), (12), 그리고 (13)을 바탕으로 PIRT 하에서의 문항 및 검사 신뢰도를 구하려면 식 (6)과 (7)을 그대로 이용하면 된다.

추가적으로, 혼합유형 문항 검사 하에서는 각각 n_1개의 이분 문항과 n_2개의 다분 문항들이 있다고 볼 수 있는데($n = n_1 + n_2$), 각 문항의 오차 변량과 진점수 변량을 구하고자 할 때에는 문항 유형에 관계없이 식 (9)부터 식 (13)까지를 그대로 사용할 수 있다. 이는 앞에서 언급된 바와 같이 이분 문항이 m=1인 특수한 상황이라고 볼 수 있기 때문이다. 유일한 차이는 각 문항 범주에 대한 반응 확률을 계산하는 데에 있어서 이분 문항을 위해서는 3PLM이 그리고 다분 문항을 위해서는 GPCM이 사용된다는 것이다.

Ⅲ. 2가지 실제 검사 자료의 문항 및 검사 신뢰도

1. 실제 자료: 영어 및 수학 검사

이 절에서는 위에서 언급된 식들을 이용하여 서로 다른 2가지 실제 자료를 대상으로 문항 및 검사 신뢰도를 구해본다. 두 자료 모두 이분 문항과 다분 문항을 함께 가지고 있는 혼합유형 문항 검사의 결과였다. 분석에 사용된 첫 번째 검사 자료(영어)는 어느 미국 중서부 대학에서 2005년도 신입생에게 실시된 배치고사의 결과 중 일부분으로, 이 논문에서 사용된 자료는 영문법에 관한 능력을 재기 위한 15개의 선택형 문항과 독해 능력을 재기 위한 25개의 선택형 문항으로 구성된 것으로, 피험자 수는 22,102명이었다. 독해 능력을 측정하기 위하여 모두 5개의 지문이 사용되었고 각기 5개의 관련 문항이 존재하였는데, 이 논문에서는 각 지문을 0점(5개 문항 모두 틀린 경우)부터 5점(5개 문항 모두 맞은 경우)까지의 여섯 개 범주를 가진 다분 문항으로 다루었다. 따라서 실제로 분석된 자료는 15개의 이분 문항과 5개의 다분 문항에 대한 피험자들의 수행 결과를 담고 있었다. 두 번째 검사 자료(수학)는 2000년도에 미국의 8학년 학생들을 대상으로

실시된 전국 수준의 수학 성취도 검사를 통해 얻어진 것으로서 4개의 선택형 문항과 5개의 서술형 문항으로 구성되었으며 피험자 수는 13,556명이었다. 수학 자료에서 5개의 서술형 문항은 각기 0, 1, 그리고 2의 세 범주로 채점되었다.

<표 1> 영어 자료의 문항 기술통계치

문항	범주별 관찰반응비율 (p_k)						평균 문항점수 ($=\sum_{k=0}^{m_i} k\,p_k$)	문항-검사간 적률상관계수	Cronbach's α if Item Deleted
	0	1	2	3	4	5			
1	.70	.30	–	–	–	–	.30	.38	.84
2	.22	.78	–	–	–	–	.78	.41	.84
3	.61	.39	–	–	–	–	.39	.40	.84
4	.40	.60	–	–	–	–	.60	.46	.84
5	.29	.71	–	–	–	–	.71	.38	.84
6	.41	.59	–	–	–	–	.59	.39	.84
7	.35	.65	–	–	–	–	.65	.39	.84
8	.72	.28	–	–	–	–	.28	.37	.84
9	.45	.55	–	–	–	–	.55	.44	.84
10	.32	.68	–	–	–	–	.68	.37	.84
11	.56	.44	–	–	–	–	.44	.35	.84
12	.40	.60	–	–	–	–	.60	.33	.84
13	.37	.63	–	–	–	–	.63	.52	.84
14	.27	.73	–	–	–	–	.73	.42	.84
15	.38	.62	–	–	–	–	.62	.39	.84
16	.02	.08	.15	.25	.30	.19	3.29	.63	.83
17	.02	.07	.15	.23	.28	.26	3.47	.62	.83
18	.01	.02	.08	.18	.34	.37	3.93	.51	.83
19	.03	.08	.16	.23	.27	.22	3.31	.57	.83
20	.02	.06	.13	.22	.27	.30	3.57	.59	.83

Reckase(1979)는 일차원성 가정을 가진 IRT 모형을 적용하기 위해서 하나의 주된 요인이 전체 점수 변량에 대해서 설명하는 비율이 적어도 20% 이상 되어야 한다고 말하였는데, 영어와 수학 자료 각각에 대하여 주성분분석을 실시한 결과 가장 큰 설명력을 갖는 주성분 변수가 전체 검사 점수 변량의 27.81%와 31.44%를 설명하는 것으로 나타나서 두 자료에 있어서 일차원성 가정 하의 문항반응모형을 적용하는 데에 문제가 없는 것으로 나타났다. 각 검사의 내적 일관성 신뢰도(Cronbach's α) 계수는 20개 문항을 가진 영어 자료의 경우 0.84이었고, 9개 문항 검사인 수학 자료의 경우 0.71이었다. 또한 전체 검사 점수의 평균(표준편차)은 영어 자료의 경우 25.83(7.37)이었고 수학 자료의 경우 5.61(3.06)이었다. <표 1>과 <표 2>에서는 이 두 자료에 대하여 먼저 CTT 하에서 제공될 수 있는 문항의 기술통계치를 각기 포함하고 있다.

<표 1>에서 볼 수 있는 바와 같이, 영어 자료의 경우 이분 문항 중에서는 문항 8이 평균 문항점수가 0.28로서 가장 어려운 문항이었고 다분 문항 중에서는 문항 16이 평균 문항점수가 3.29로서 가장 낮아서 제일 어려운 문항인 것으로 나타났다. 문항과 검사간 상관계수를 구한 결과, 이분 문항 중에서는 문항 13(r=0.52)이 그리고 다분 문항 중에서는 문항 16(r=0.63)이 가장 큰 변별력을 보여 주었다. 각 문항이 제외되었을 때의 전체 검사의 신뢰도를 보면 0.84 혹은 0.83으로서 문항 간에 거의 차이가 없었으며, 다시 말해서 전체 검사 신뢰도에 미치는 각 문항의 공헌도는 거의 비슷하였다.

<표 2> 수학 자료의 문항 기술통계치

문항	범주별 반응비율			평균 문항점수 ($= \sum_{k=0}^{m_i} k\, p_k$)	문항-검사간 적률상관계수	Cronbach's α if Item Deleted
	0	1	2			
1	.29	.71	-	.71	.27	.71
2	.46	.54	-	.54	.45	.68
3	.72	.28	-	.28	.37	.69
4	.64	.36	-	.36	.23	.71
5	.10	.68	.22	1.12	.51	.67
6	.28	.47	.25	.97	.38	.69
7	.46	.15	.39	.93	.45	.68
8	.66	.22	.11	.45	.49	.67
9	.84	.05	.10	.26	.39	.69

<표 2>는 수학 자료의 문항 기술통계치를 소개하고 있다. 우선 이분 문항 중에서는 문항 3이 가장 어렵고 문항 1이 가장 쉬운 문항인 것으로 나타났고 다분 문항 중에서는 문항 9가 가장 어렵고 문항 5가 가장 쉬운 것으로 나타났다. 특히 문항 9의 경우 0, 1, 그리고 2의 범주를 가진 다분 문항임에도 불구하고 평균 문항점수가 0.26에 지나지 않아서 매우 어려운 문항인 것으로 보였다. 문항과 검사 간 상관계수를 보았을 때, 이분 문항 중에서는 문항 2(r=0.45)가 그리고 다분 문항 중에서는 문항 5(r=0.51)가 가장 큰 변별력을 가지는 것으로 나타났다. 전체 9개 문항 중에서 문항 4(r=0.23)가 가장 낮은 변별력을 보여 주었다. 전체 검사 신뢰도에 대한 각 문항의 공헌도를 보면, 문항 1과 4의 경우 각기 검사에서 제외된다고 하더라도 전체 신뢰도에 거의 변화를 주지 않는 것으로 나타나서 별다른 공헌도가 없는 것으로 나타났다. 하지만 문항 5와 8의 경우에는 각기 제외될 경우 전체 신뢰도가 0.67로 나타나서 검사 신뢰도에 미치는 영향이 가장 큰 것으로 나타났다.

다음으로 PARSCALE(Muraki & Bock, 2003) 프로그램을 가지고 각 검사에서의 문항 모수를 추정하였는데, <표 3>과 <표 4>에서는 이러한 결과가 보고되었다. 혼합유형 문항 검사 자료를 다루기 위한 PARSCALE 코드는 Kim & Lee(2004)를 참조하여 생성하였다. <표 3>에서 볼 수 있는 바와 같이 영어 자료의 경우, 이분 문항 중에서는 문항 13이 그리고 다분 문항 중에서는 문항 16이 가장 변별력 있는 문항으로 나타났다. 또한 문항 곤란도를 볼 때 이분 문항 중에서는 문항 2가 가장 쉽고 문항 8이 가장 어려운 문항이었으며, 다분 문항 중에서는 문항 16이 가장 어렵고 문항 18이 가장 쉬운 문항이었다. <표 4>의 수학 자료 문항 모수 추정치를 보면, 이분 문항 중에서는 문항 2가 그리고 다분 문항 중에서는 문항 5가 가장 변별력이 큰 문항이었다. 문항 곤란도를 보면, 이분 문항 중에서는 문항 1이 그리고 다분 문항 중에서는 문항 5가 가장 쉬운 문항이었다. τ 추정치를 보면 문항 5, 6, 그리고 8에서는 $\hat{\tau}_1$ 보다 $\hat{\tau}_2$ 가 작지만, 문항 7과 9에서는 그 반대임을 알 수 있다. 이는 범주 위치 모수가 범주 값이 커짐에 따라 일정한 순서를 따르는 Samejima(1969)의 등급반응모형(GRM: graded response model)과는 달리, GPCM 하에서는 이러한 가정이 존재하지 않기 때문이다.

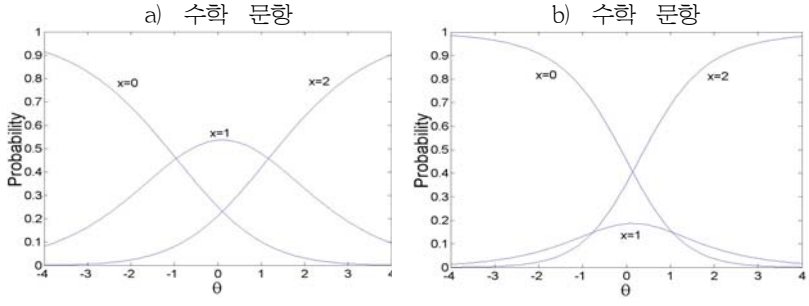
<표 3> 영어 자료에서의 문항모수 추정치

문항		문항모수 추정치							
		α_i	β_i	γ_i	τ_1	τ_2	τ_3	τ_4	τ_5
이분 ($m=1$)	1	1.25	.97	.03	–	–	–	–	–
	2	1.46	−1.02	.15	–	–	–	–	–
	3	1.32	.61	.06	–	–	–	–	–
	4	1.53	−.24	.09	–	–	–	–	–
	5	1.46	−.36	.28	–	–	–	–	–
	6	1.54	.10	.23	–	–	–	–	–
	7	1.18	−.52	.10	–	–	–	–	–
	8	1.39	1.09	.04	–	–	–	–	–
	9	1.49	.00	.11	–	–	–	–	–
	10	1.40	−.25	.27	–	–	–	–	–
	11	1.27	.66	.14	–	–	–	–	–
	12	1.04	−.05	.19	–	–	–	–	–
	13	1.75	−.42	.05	–	–	–	–	–
	14	1.70	−.48	.27	–	–	–	–	–
	15	1.01	−.62	.00	–	–	–	–	–
다분 ($m=5$)	16	.98	−.82	–	1.86	.70	.12	−.70	−1.97
	17	.94	−1.08	–	1.86	.76	−.11	−.83	−1.68
	18	.73	−1.81	–	1.70	.90	.11	−.71	−2.01
	19	.73	−.95	–	1.83	.81	−.07	−.79	−1.78
	20	.79	−1.18	–	1.48	.84	−.03	−.81	−1.48

이러한 GPCM의 특성에 대한 이해를 돕기 위하여 문항 6과 7의 ICCC가 [그림 1]에 제시되었다. 흔히 $\beta_i - \tau_{ik}$는 단계(step) 모수라고 불리는데 이는 문항 i의 범주 $k-1$ 과 k간의 상대적 곤란도를 의미한다. 문항 6에서 범주의 값 혹은 문항 점수 $x=1$과 $x=2$를 위한 단계 모수는 각각 -0.96(=0.10-1.06)과 1.16(=0.10+1.06)이었다. 다시 말해서 문항 점수 1에서 2를 획득할 때의 상대적 곤란도가 문항 점수 0에서 1을 획득할 때의 곤란도 보다 크다. 반대로 문항 7에서는 문항 점수 $x=1$과 $x=2$를 위한 단계 모수가 각각 1.05(=0.14+0.91)과 -0.77(=0.14-0.91)이 된다. 이는 문항 점수 1에서 2를 획득할 때의 상대적 곤란도가 문항 점수 0에서 1을 획득할 때보다 더 작다는 것이다. 이는 다시 말해서 어떤 피험자의 문항 수행에 대해서 점수가 부여될 때 0에서 1로 문항 점수가 바뀌는 것은 힘들지만 1에서 2로 가는 것은 상대적으로 수월하다는 것을 의미한다. 결과적으로 [그림 1-b]에서 볼 수 있는 바와 같이 ICCC 상에서 $x=1$ 에 해당하는 범주특성곡선은 θ척도 위에서 단 한 번도 최대가 되지 못하였다. 따라서 대부분의 피험자들은 이 문항에 대해서 0점 아니면 2점을 얻게 될 것임을 알 수 있다. 실제로 13,556명의 피험자 중에서 문항 점수 0, 1, 그리고 2점을 얻은 비율은 <표 2>에서 확인할 수 있는 바와 같이, 문항 6의 경우 28%, 47%, 그리고 25%이었지만 문항 7의 경우는 46%, 15%, 그리고 39%이었다.

<표 4> 수학 자료에서의 문항모수 추정치

문항		문항모수 추정치				
		α_i	β_i	γ_i	τ_1	τ_2
이분 ($m=1$)	1	.91	−.85	.16	−	−
	2	1.97	.05	.11	−	−
	3	1.47	1.10	.05	−	−
	4	.90	1.41	.15	−	−
다분 ($m=2$)	5	1.85	−.35	−	1.38	−1.38
	6	.79	.10	−	1.06	−1.06
	7	.85	.14	−	−.91	.91
	8	1.31	1.15	−	.25	−.25
	9	.95	1.65	−	−1.46	1.46



[그림 1] 수학 문항 6과 7의 ICCC

2. 추정된 문항 및 검사 신뢰도

<표 3>과 <표 4>에서 주어진 문항 모수 추정치를 바탕으로 영어 및 수학 자료에서 IRT에 기반을 둔 문항 및 검사 신뢰도가 추정되었고, 이는 <표 5>와 <표 6>에 각각 제시되었다. 이러한 계산을 위하여 PARSCALE의 결과 파일에서 구해진 문항 모수 추정치를 사용하였고, 역시 같은 파일에서 제공되는 θ의 사후분포를 확률밀도분포인 $\phi(\theta)$로서 사용하였다. 이 사후분포는 20개의 구적점들(quadrature points)을 가진 이산분포 형태였으며 이를 이용하여 식 (2), (3), (5), (9), (11), 그리고 (13) 등에 대한 수치적분(numerical integration)을 실시하였다.

<표 5> 영어 자료에서의 문항 및 검사 신뢰도

문항	문항 오차변량 및 진변량, 그리고 신뢰도 추정치					
	$\hat{\sigma}^2(e_i)$	$\hat{\sigma}^2(\tau_i)$	$\hat{\rho}_{ii}$	if n=20*	if n=40*	if n=60*
1	.17	.04	.19	.82	.90	.93
2	.14	.03	.18	.81	.90	.93
3	.19	.05	.21	.84	.91	.94
4	.18	.06	.25	.87	.93	.95
5	.17	.04	.19	.82	.90	.93
6	.20	.05	.20	.83	.91	.94
7	.19	.04	.17	.80	.89	.92
8	.16	.04	.20	.83	.91	.94
9	.19	.06	.24	.86	.93	.95
10	.18	.04	.18	.81	.90	.93
11	.21	.04	.16	.79	.88	.92
12	.21	.03	.13	.75	.86	.90
13	.16	.07	.30	.90	.94	.96
14	.16	.04	.20	.83	.91	.94
15	.19	.04	.17	.80	.89	.92
16	.89	.78	.47	.95	.97	.98
17	.91	.75	.45	.94	.97	.98
18	.80	.36	.31	.90	.95	.96
19	1.12	.67	.37	.92	.96	.97
20	1.03	.68	.40	.93	.96	.98
검사 신뢰도	$\hat{\rho}_{xx} = \frac{\hat{\sigma}^2{}_{\tau}}{\hat{\sigma}^2{}_{\tau} + \hat{\sigma}^2{}_{e}} = \frac{49.49}{49.49 + 7.45} = 0.87$					

* 해당 문항과 동일한 신뢰도의 n개 문항을 가진 가상 검사의 신뢰도를 Spearman-Brown 공식으로 계산.

<표 5>에 제시된 영어 자료에서의 문항 및 검사 신뢰도 추정치를 보면, 우선 전체 검사의 신뢰도가 0.87로 추정되었는데 이는 자료를 통해 추정된 Cronbach's α(0.84)에 비해 비슷하지만 약간 큰 값이었다. 전반적으로 다분 문항들이 이분 문항들에 비하여 높은 문항 신뢰도를 가지는

것으로 나타났고, 이분 문항 중에서는 문항 13이 그리고 다분 문항 중에서는 문항 16이 가장 높은 문항 신뢰도를 가졌다. 널리 알려져 있다시피 검사의 신뢰도는 문항의 수가 많아질수록 커지게 되는데, 한 문항에 대한 신뢰도를 의미하는 ρ_{ii}를 해석할 때 그 준거가 모호한 문제가 있다. 즉 CTT에서 검사 신뢰도를 해석할 때 표준화검사의 경우에는 대략 0.8 내지 0.9 그리고 교실 수준의 검사에서는 0.6 이상의 신뢰도를 요구하는 것으로 알려져 있지만, 문항 신뢰도 값에 대해서는 이러한 기준이 정해져 있지 않다. 이를 해결하기 위하여 <표 5>의 다섯 번째부터 일곱 번째 열에서는 각종 학업성취도 검사의 문항수가 대략 20개 - 60개 사이인 점을 감안하여 Spearman-Brown 공식을 이용하여 해당 문항과 완전히 동등한 문항들이 각각 20, 40, 그리고 60개인 경우의 가상 검사가 가지게 되는 신뢰도를 계산해 보았다. 각 문항별로 분석해 보았을 때, 같은 문항 신뢰도를 가진 20 문항으로 이루어진 가상의 검사들을 보면 여전히 검사 신뢰도가 0.80에 미치지 못하는 경우가 존재하였다. 하지만, 40개 문항 이상부터는 모든 문항에서 0.86이상의 검사 신뢰도가 도출될 것으로 추정되었다. 예를 들어, 문항 12의 문항 신뢰도는 0.13이었는데 이러한 문항이 20개 있는 가상 검사의 신뢰도는 0.75였지만 40개가 있는 검사의 경우 신뢰도가 0.86에 달할 것으로 예상되었다. 이러한 정보를 바탕으로 Cronbach's α if item deleted를 사용하듯이 여러 예비 문항 중 필요 문항수만큼 최고의 문항들을 선출해 내는 데에 활용해 낼 수 있을 것이며, 각 문항의 신뢰도를 다른 문항 신뢰도와의 상대적 비교가 아닌 절대적 기준에 의해 판단하는 것도 가능할 것이다. 또한, 이러한 신뢰도 계산이 실제 검사의 실시 없이도 계산할 수 있는 정보라는 점에서 그 유용성이 배가된다고 볼 수 있다 .

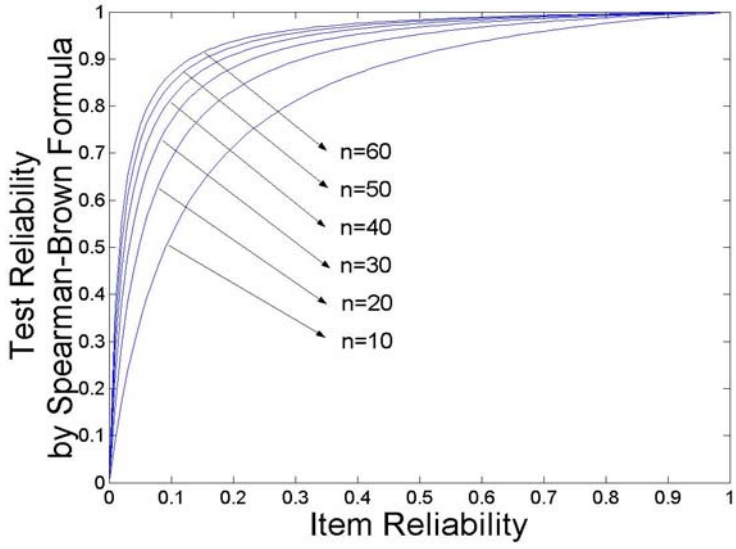
<표 6> 수학 자료에서의 문항 및 검사 신뢰도

문항	문항 오차변량 및 진변량, 그리고 신뢰도 추정치					
	$\hat{\sigma}^2(e_i)$	$\hat{\sigma}^2(\tau_i)$	$\hat{\rho}_{ii}$	if n=20*	if n=40*	if n=60*
1	.18	.02	.10	.69	.82	.87
2	.17	.08	.32	.90	.95	.97
3	.16	.04	.20	.83	.91	.94
4	.21	.02	.09	.66	.80	.86
5	.19	.12	.39	.93	.96	.97
6	.42	.11	.21	.84	.91	.94
7	.58	.26	.31	.90	.95	.96
8	.30	.18	.38	.92	.96	.97
9	.29	.10	.26	.88	.93	.95
검사 신뢰도	$\hat{\rho}_{xx} = \frac{\hat{\sigma}^2{}_{\tau}}{\hat{\sigma}^2{}_{\tau+}\hat{\sigma}^2{}_{e}} = \frac{6.84}{6.84+2.50} = 0.73$					

* 해당 문항과 동일한 신뢰도의 n개 문항을 가진 가상 검사의 신뢰도를 Spearman-Brown 공식으로 계산.

수학 자료의 경우 이분 문항의 경우 문항 4의 신뢰도가 0.09로서 가장 낮았고 문항 2의 신뢰도가 가장 높았다. 다분 문항 중에서는 문항 6의 신뢰도가 0.21이었지만 문항 5의 경우 0.39에 달하여 매우 높은 문항신뢰도를 보여주었다. 영어 자료에서와 마찬가지로, 같은 문항이 20, 40, 그리고 60개 있는 경우의 가상 검사의 신뢰도를 구해본 결과 40문항 이상일 때 모든 검사 신뢰도는 0.80 이상이었다. 만약 20문항을 가지고 검사를 만든다면, 문항 1과 4 같은 신뢰도를 가지는 문항들로 검사가 구성될 경우 그 검사 신뢰도가 0.70보다 작게 될 것임을 예상할 수 있다. 이 논문에서 다루어진 문항 및 검사 신뢰도의 추정을 위하여 MATLAB 프로그램으로 직접 작성된 코드가 사용되었는데, 본문 뒤의 <부록>에는 수학 자료를 위하여 적용된 코드가 첨부되어 있다.

[그림 2]에서는 이와 같은 문항 신뢰도의 해석을 돕기 위하여, 문항 신뢰도가 0에서 1까지 변화할 때 가상의 검사들(n=10, n=20, n=30, n=40, n=50, 그리고 n=60)이 가질 것으로 예상되는 신뢰도를 Spearman-Brown 공식을 활용하여 그래프로 나타내 보았다. 예를 들어, 검사 제작자가 신뢰도가 0.80이상인 검사를 만들고자 할 때 IRT 모수의 함수로서 구해진 문항 신뢰도를 가지고 적합한 문항을 고르고자 하는 상황을 생각해 볼 수 있다. 그 검사의 총 문항 수가 10개라면 아래의 [그림 2]에서 제공되는 정보를 사용해서 볼 때 한 문항의 신뢰도는 대략 0.28 내지 0.29 정도가 되어야 할 것이다. 반면 총 문항 수가 20개로 늘어나면 필요한 문항 신뢰도는 0.17 정도가 될 것이며, 더 나아가서 60개라면 0.07정도의 문항 신뢰도로도 충분할 것이다.



[그림 2] 문항 신뢰도와 검사 신뢰도의 관계

Ⅳ. 요약 및 논의

흔히 CTT의 신뢰도에 대응되는 개념으로서 IRT에서의 정보함수나 추정의 오차변량이 언급되는 경향이 있다. 정보함수는 피험자의 능력 수준에 따라서 검사를 통해서 얻을 수 있는 능력 추정의 상대적 정확도에 대한 파악을 가능하게 해주는 장점이 있다. 하지만, 과연 어느 정도의 정보함수 값이 만족할만한 수준의 정확한 능력의 측정을 의미하는 지에 대해서는 단일한 판단의 기준을 확립하기가 어렵다. 따라서 측정의 정확성에 대해서 논의하기 위해서는 어떠한 기준점(benchmark)이 필요하게 되는데, 신뢰도 계수에서 이용되는 관찰점수의 변량은 이처럼 오차점수의 변량의 정도를 일정한 기준을 가지고 판단하게 해 주는 역할을 한다(Kane, 1996). 따라서 IRT를 통해서 검사 결과를 분석할 때, 능력을 재는 정확성의 정도에 관한 다양한 지표를 얻기 위해서는 문항 및 검사의 정보함수와 능력추정의 표준오차에 더불어서 문항 및 검사의 신뢰도를 파악하는 것이 여전히 매우 중요하다고 말할 수 있다(Samejima, 1994). 본 논문에서는 혼합유형 문항 검사 자료에 대하여 문항반응모형 모수의 함수로써 신뢰도를 계산하는 방법을 소개함으로써, 기존의 내적일관성 지수와 같은 신뢰도를 구할 수 없는 상황 즉 다수의 예비 검사집이 제작되었거나 CAT가 실시되는 경우에 있어서도 편리하게 검사 혹은 개별 문항의 신뢰도를 추정할 수 있음을 보였다.

<표 5>와 <표 6>의 문항 및 검사 신뢰도를 계산하기 위하여, 앞에서 밝힌 바와 마찬가지로, 각 자료에서의 모수 추정 과정을 통해 구해진 θ의 사후분포를 확률밀도분포인 $\phi(\theta)$ 에 대응하는 개념으로 사용하였다. 이는 실제 자료가 존재하는 경우이기 때문에 가능하였지만, 만약 다수의 예비 검사집에 대한 신뢰도를 실제 검사의 실시 없이 추정해 보는 상황 하에서는 이러한 경험적 사후분포를 구하기가 곤란하게 된다. 이런 경우, 주변최대가능도(MML: marginal maximum likelihood) 방법으로 문항 모수를 추정할 때 θ에 대하여 흔히 적용되는 사전분포인 N(0,1)이 $\phi(\theta)$로서 사용되는 것이 하나의 대안이 될 수 있을 것이다. 위의 영어와 수학 자료에 대하여 사후분포 대신에 N(0,1)을 가지고 검사 신뢰도를 추정해 보면 각각 0.87과 0.73로서 소수점 셋째 자리 이하에서 약간의 차이를 보였을 뿐 별 차이가 없었으며 문항 신뢰도 역시 매우 작은 차이가 존재할 뿐이었다. 따라서 큰 표집을 사용하는 경우, 피험자 능력의 사후분포를 통하여 계산된 문항 및 검사 신뢰도와 N(0,1)을 사용하여 구한 신뢰도와는 큰 차이가 없는 것으로 보였다. 하지만 보다 확실하게 능력 분포가 신뢰도 추정에 미치는 영향을 검토하기 위해서는 다양한 능력분포, 문항의 수, 피험자의 수, 문항 모수의 크기 등등의 여러 가지 상황을 함께 고려하는 시뮬레이션 연구가 수행될 필요가 있다.

이하에서는 크게 두 가지 사항과 관련하여 본 연구의 시사점을 검토해 보았다. 하나는 IRT에 기반하여 신뢰도를 추정하는 다른 접근 방법을 간략히 소개하고 상호 차이점에 대하여 논의하는 것이며, 또 다른 하나는 현재 가장 널리 쓰이는 신뢰도 산출 방법인 Cronbach's α와의 이론적 비교이다.

IRT 분야에서는 이 논문에서 소개된 신뢰도 추정법 외에도 문항 모수 추정치들의 함수로서 신뢰도를 계산하기 위한 시도가 있어 왔는데, Green et al. (1984)이 제안한 주변적 신뢰도(marginal reliability) 계수를 대표적인 예로 들 수 있다. 주변적 신뢰도는 θ 추정의 오차 변량(error variance of estimation)의 평균을 적분을 통하여 구한 다음에 피험자 모수로서 추정된 θ 값들의 변량을 함께 이용하여 다음과 같이 계산할 수 있다:

$$\text{주변적 신뢰도} = \frac{\text{추정된 } \theta\text{의 변량} - \theta \text{ 추정의 오차 변량 평균}}{\text{추정된 } \theta\text{의 변량}} \qquad (14)$$

신뢰도는 엄밀히 말해 검사 자체의 특성이라기보다는 검사 결과 산출된 점수들의 정확도와 관련된 특성을 나타내는데(Brennan, 2001), 이런 관점에서 위의 주변적 신뢰도는 검사 원점수의 신뢰도에 관한 것이라기보다 추정된 θ 점수의 신뢰도라고 볼 수 있다. 검사 점수의 비선형적 척도 변환이 이루어질 때 새로운 척도점수 하에서 계산되는 신뢰도는 달라질 수 있기 때문에(Kolen, Hanson, & Brennan, 1992), 식 (14)를 통해 구해진 주변적 신뢰도는 앞의 식 (7)을 통해서 구해진 검사 신뢰도와 다소 차이가 있을 것이다. CAT 상황에서 주변적 신뢰도를 이용하고자 할 때, 흔히 '추정된 θ의 변량'을 1로 볼 것이 추천되는데(Thissen, 1990), 이는 앞에서 $\phi(\theta)$를 위하여 N(0,1)을 사용하는 것이 타당할 것인가와 비슷한 문제로 볼 수 있다. 어떠한 신뢰도를 어떠한 상황 하에서 사용하는 것이 보다 바람직한가에 대한 정보를 얻기 위해서는 보다 체계적인 후속 연구가 필요할 것으로 보인다.

또한 Kolen, Zeng, & Hanson(1996)의 연구에서는 Lord & Wingersky(1984)의 재귀 공식(recursive formula)를 이용하여 조건적 관찰점수분포, 즉 $\Pr(X|\theta)$를 구하고 이를 다시 문항반응모형의 맥락 속에서 오차점수 변량을 추정하는 데에 사용할 수 있음을 제안하였다. 이러한 접근방법 속에서는 원점수가 선형적 혹은 비선형적으로 변환된 척도점수(scale score)가 사용될 때에도 손쉽게 관련 검사 신뢰도를 추정할 수 있다는 장점이 있다. 하지만, 이들의 연구에서는 본 연구에서 소개된 문항 수준의 신뢰도를 구하는 것은 논의하지 않았고, 또한 관찰점수 변량을 구하기 위하여 실제 검사를 실시한 결과 얻은 자료가 요구된다는 점에서 신뢰도 추정을 위하여 문항모수 추정치만을 요구하는 본 연구에서의 방법과 차이가 있다. 주지할 점은 본 논문에서 소

개된 Dimitrov(2003)의 접근방법이든, 주변적 신뢰도이든, Kolen 등의 방법이든 모두 적용된 문항반응모형이 주어진 자료를 잘 적합한다는 전제 위에서 그 의의를 가진다는 점이다.

앞에서 언급된 것과 같이, 영어의 경우 22,102명의 자료에서 계산된 Cronbach's α가 0.84이었고 수학의 경우 13,556명의 자료에서 계산된 Cronbach's α가 0.71이었다. 이들은 IRT 모수의 함수로서 추정된 검사 신뢰도(각각 0.87과 0.73)에 비해서 매우 가까우면서도 약간씩 작은 값이었는데, Kolen, Zeng, & Hanson (1996)의 연구에서도 유사한 사례가 보고된 바 있다. 이러한 현상이 일반적인 것인지 혹은 영어와 수학 자료에 특정된 것인지는 보다 심층적인 논의가 필요할 것으로 보이는데, 우선 Lee, Brennan, & Kolen(2000)에서 그 설명의 단서가 되는 연구 결과를 찾아볼 수 있다. 그들의 시뮬레이션 연구에 따르면 같은 자료에 대하여 문항반응모형을 가지고 계산된 능력 수준에 따른 조건적 측정의 표준오차(conditional standard errors of measurement)가 강한 진점수 이론 하의 이항분포모형이나 혼합이항분포모형에 의해서 계산된 값보다 작은 경향이 있었다. 이러한 현상이 나타나는 주된 이유는 문항반응모형이 고정평형검사(fixed parallel forms)를 가정하는 반면에 CTT에서는 무선평형검사(random parallel forms)를 가정하기 때문으로 설명되었다. 다시 말해서, 다른 모형들과는 달리 문항반응모형이 말하는 측정의 오차에는 문항 전집으로부터의 문항 내용 등과 관련된 표집의 오차가 고려되지 않는다는 것이다.

Cronbach's α와 이 연구에서 제시된 IRT를 이용하는 신뢰도 계수의 차이점은 또한 문항 수준의 측정모형(measurement model)의 관점에서 논의될 수 있다. 주지하다시피 Cronbach's α는 여러 가지 측정 모형들 중에서 유사진점수동등모형(essentially tau equivalent model)에 바탕을 둔 신뢰도 추정 방법이다(Novick & Lewis, 1967; Graham, 2006). 이 측정모형은 각 문항의 관찰점수에 영향을 주는 상수로서의 문항곤란도에 있어서 차이가 있을 수 있다고 가정한다. 하지만 일반공동모형(congeneric model)에서처럼 한 개인의 진점수가 각 문항 관찰 점수에 기여하는 정도가 선형적으로 다를 수 있다고까지는 가정하지 않기 때문에, 검사 자료들이 일반공동모형으로 보다 잘 설명될 시에 Cronbach's α는 신뢰도를 과소추정하게 된다(Novick & Lewis, 1967; McDonald, 1999). 혼합유형 문항 검사의 경우 문항 점수의 특성과 크기가 상호 다르게 됨에 따라서 진점수동등성(tau equivalency)을 유지할 수 없는 경우가 많기 때문에 (Qualls, 1995), Cronbach's α의 적용과 해석은 주의 깊게 이루어져야 한다. 본 논문에서 제시된 신뢰도 추정 기법은 유사진점수동등모형(essentially tau equivalent model)의 제약을 받지 않고 다만 검사 문항들이 하나의 잠재변수를 측정하고 있다는 단일요인모형(single factor model)의 가정만을 가지기 때문에, 일차원성을 충족시키는 혼합유형 문항 검사 결과의 신뢰도 추정에 적용할 때 측정모형 이론의 관점에서 보다 안전한 방법이라고 생각된다.

참고문헌

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*. (Eds. F. M. Lord and M. R. Novick). Reading: Addison-Wesley, pp. 397-472.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*, 295-317.

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement*, *27*, 440-458.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement*, *66*, 930-944.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347-360.

Kane, M. (1996). The precision of measurements. *Applied Measurement in Education*, *9*, 355-379.

Kim, S., & Lee, W. (2004). *IRT scale linking methods for mixed-format tests* (ACT Research Report 2004-5). Iowa City, IA: ACT, Inc.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard error of measurement for scale scores, *Journal of Educational Measurement, 29*, 285-307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*, 129-140.

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37*, 1-20.

Lord, F. M., & Novick, M. R. (1968). *Statistical theory of mental test scores*. Reading, MA: Addison-Wesley.

May, K., & Nicewander, W. A. (1994). Reliability and information functions for percentile ranks. *Journal of Educational Measurement*, *31*, 313-325.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Lawrence Erlbaum.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muraki, E. & Bock, R. D. (2003). *PARSCALE (version 4.1): IRT Item Analysis and Test Scoring for Rating-Scale Data*. Chicago, IL: Scientific Software, Inc.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1-13.

Raju, N. S., Price, L. R., Oshima, T. C. & Nering, M. L. (in press). Standardized conditional SEM: a case for conditional reliability, *Applied Psychological Mea- surement*.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*, 3-32.

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education*, *8*, 111-120.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *17*.

Samejima, F. (1994). Estimation of reliability coefficient using the test information function and its modifications. *Applied Psychological Measurement*, *18*, 229-244.

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 161-186). Hillsdale, NJ: Lawrence Erlbaum.

* 논문접수 2011년 1월 18일 / 1차 심사 2011년 2월 11일 / 2차 심사 2011년 3월 7일 / 게재승인 2011년 3월 9일

* 강태훈(姜兌勳, Kang, TaeHoon) : 서울대학교 교육학과를 졸업하고 동 대학원 교육학과에서 석사학위를 취득 후, 미국 위스콘신 대학교(UW-Madison) 교육심리학과에서 양적방법론-교육측정 분야 박사(Ph.D.) 학위를 취득하였음. 현재 성신여자대학교 교육학과 조교수로 재직 중이며 주요 연구 분야는 문항반응이론과 관련하여 검사 동등화, 다차원 문항반응이론, 통계적 모형의 선택 및 적합도 검증 그리고 컴퓨터 게임을 통한 인지 능력의 평가 등임.
* E-mail : taehoonkang@gmail.com, * homepage : http://irt.com.ne.kr,

부 록

IRT 모수의 함수로서 문항 및 검사 신뢰도를 구하기 위한 MATLAB 코드

```
-----------------------------------------------------------------------
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%          Reliability as a function of item and theta parameter estimates :          %%%
%%%                    Mixed-Format Test Data of Math Achievement                    %%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Math data (9 items and 20 Quadrature points)
ndi=4; npi=5; load 'itempara.txt'; load 'abilpara.txt';
lam=itempara(1:ndi,1); bet=itempara(1:ndi,2); gue=itempara(1:ndi,3);
gquad=abilpara(:,1); densi=abilpara(:,2); K=ndi+npi; n=20;

% Dichotomous items: calculating Pi(theta) by the 3PLM
for i=1:ndi
    for j=1:n
        logit(j,i) = lam(i) * (gquad(j)-bet(i));
        par1(j,i) = gue(i)+(1-gue(i))*(1/(1+exp(-logit(j,i))));
    end
end
% Polyotmous items: calculating Pi(theta) by the GPCM
a=itempara(ndi+1:K,1); b=itempara(ndi+1:K,2); tau2=itempara(ndi+1:K,3);
tau3=itempara(ndi+1:K,4);
ncat=3; par2=zeros(n,npi,ncat); tt=zeros(npi,n,ncat); denom=zeros(npi,n);
for t=1:n
        for i=1:npi
        tt(t,i,1) = 1;
        tt(t,i,2) = exp(a(i)*(gquad(t)-b(i)+tau2(i)));
        tt(t,i,3) = exp(a(i)*(gquad(t)-b(i)+tau2(i) + gquad(t)-b(i)+tau3(i)));
        denom(t,i) = 1 + tt(t,i,2) + tt(t,i,3);
    end
end
for t=1:n
    for i=1:npi
        for w=1:ncat
            par2(t,i,w)=tt(t,i,w)/denom(t,i);
        end
    end
end
```

```
% dichotomous and polytomous together
par=zeros(n,K,ncat); par(:,1:ndi, 1)=1-par1; par(:,1:ndi, 2)=par1; par(:,ndi+1:K,:)=par2;

% item error and true-score variances
irelia=zeros(K,1); iev=zeros(K,1); tsv=zeros(K,1); imsit1=zeros(K,1); imsit2=zeros(K,1);
for i=1:K
    for t=1:n
        kk=0; k=0;
        for z=1:ncat
            kk = kk + (z-1)^2 * par(t,i,z);
            k = k + (z-1) * par(t,i,z);
        end
        iev(i)=iev(i) + (kk - k^2)*densi(t);
        imsit1(i)=imsit1(i) + k^2 * densi(t);
        imsit2(i)=imsit2(i) + k * densi(t);
    end
end
for i=1:K
    tsv(i)=imsit1(i) - imsit2(i)^2;
end
tsv = roundn(tsv,-2); iev = roundn(iev,-2);
for i=1:K
    irelia(i)=tsv(i) / (tsv(i) + iev(i));
end

% error and true score variances for Test Reliability
TTV=0; TEV=sum(iev); iTTV1=0; iTTV2=0;
R = zeros(n,K);
for i=1:K
    for t=1:n
        for z=1:ncat
            R(t,i) = R(t,i) + (z-1) * par(t,i,z);
        end
    end
end
TCC=sum(R');

for t=1:n
    iTTV1 = iTTV1 + TCC(t)^2 * densi(t);
    iTTV2 = iTTV2 + TCC(t) * densi(t);
end
TTV = iTTV1 - iTTV2^2;

% item and test reliability
itemreli=roundn(irelia,-2)
testreli = (TTV ) / (TTV + TEV)
```

itempara.txt

0.9059	−0.8505	0.1588	0
1.9667	0.0450	0.1118	0
1.4660	1.0962	0.0496	0
0.9006	1.4086	0.1462	0
1.8462	−0.3481	1.3782	−1.3782
0.7924	0.1047	1.0628	−1.0628
0.8531	0.1418	−0.9100	0.9100
1.3121	1.1521	0.2491	−0.2491
0.9499	1.6524	−1.4571	1.4571

abilpara.txt

−4.0000	0.0001
−3.5790	0.0003
−3.1580	0.0012
−2.7370	0.0041
−2.3160	0.0116
−1.8950	0.0278
−1.4740	0.0566
−1.0530	0.0970
−0.6316	0.1381
−0.2105	0.1632
0.2105	0.1624
0.6316	0.1378
1.0530	0.0982
1.4740	0.0576
1.8950	0.0279
2.3160	0.0112
2.7370	0.0037
3.1580	0.0011
3.5790	0.0003
4.0000	0.0001

Abstract

Estimating the Item and Test Reliability as Functions of the IRT parameters

Kang, TaeHoon*

This article introduces a technique to estimate item and test reliability for a psychological and educational test as a function of the item response model parameters. Dimitrov(2003) suggested a framework to calculate item and test reliability coefficients for a dichotomous item test as an application of IRT. This study generalized his method into the application for polytomous and mixed-format item tests. Two actual mixed-format test data sets were used to illustrate the use of the reliability estimation technique. The item level reliabilities seemed to be useful in the context of developing and evaluating test items. The estimated reliability based on IRT tended a little bit larger than Cornbach's α coefficient. The relationship and difference between the two reliability estimation methods were further discussed.

Key words: Item Response Theory, Mixed-Format Item Test, Item Reliability, Test Reliability

* Assistant Professor, Sungshin Women's University