교육평가연구
Journal of Educational Evaluation
2018, 제31권 제3호, 609-630
http://dx.doi.org/10.31158/JEEV.2018.31.3.609

표본크기가 S-X^2 문항 적합도 지수의 제1종 오류에 미치는 영향*

강 태 훈**
성신여자대학교

본 연구에서는 카이제곱 분포를 따르는 S-X^2 문항 적합도 지수가 제1종 오류에 있어서 표본크기의 영향을 어느 정도 받는지에 관하여 모의실험을 통해 살펴보았다. 추가적으로, 문항반응이론 분야에서 전통적으로 활용되어 온 문항 적합도 지수인 Q_1-χ^2의 수행도 함께 조사하였다. 모의실험에서는 여러 이분 문항반응모형과 다양한 수준의 문항 수를 고려하여, 제1종 오류와 통계적 검정력 측면에서 문항 적합도 검정이 각 조건에 따라 어떻게 이루어지는지 확인하였다. 연구 결과, 검사 자료 분석을 위하여 문항반응이론을 활용하고자 하는 연구자는 BILOG-MG와 같이 일반적으로 널리 쓰이는 프로그램에서 제공하는 문항 적합도 지수보다 S-X^2 지수를 사용하는 것이 훨씬 더 바람직하다고 볼 수 있었다. 또한 표본크기가 매우 클 때에도 S-X^2 지수를 사용한다면 무선적 표집 절차 없이 모든 검사 응답 자료를 활용하여 문항 적합도 검정을 수행할 수 있는 것으로 나타났다.

주제어 : 문항반응이론, 표본크기, 문항 적합도 지수, 경험적 제1종 오류, 통계적 검정력

* 이 논문은 2016년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.
** 교신저자 : 강태훈, 성신여자대학교, taehoonkang@gmail.com

Ⅰ. 서 론

교육 및 심리 검사의 결과 자료를 문항반응이론(item response theory, IRT)을 통하여 분석하고자 할 때, 기본적인 IRT 가정의 검증에 더하여 적용하고자 하는 문항반응모형이 주어진 자료를 적절히 요약 및 설명하고 있는지를 확인하는 모형 적합도 검정이 필수적으로 수행되어야 한다(설현수, 2014; 임미경 & 허선, 2007; McKinley & Mills, 1985). 주지하다시피 IRT에서의 자료 분석은 하나의 심리측정학적 모형을 상정할 경우 이러한 모형이 자료를 제대로 설명할 것이라는 기대 하에서 문항 모수 및 피험자 능력 모수를 추정하게 된다. 따라서 관찰점수가 진점수 및 오차점수로 구성된다고 보는 간단한 형태의 모형만을 갖는 고전검사이론에서와 달리, 자료 분석에 있어서 모형과 자료의 적합도를 살피는 과정이 보다 중요시 될 수밖에 없다. 이에 따라 수많은 IRT 관련 참고문헌에서는 문항 수준에서의 모형 적합도를 살펴보기 위한 다양한 방법을 소개하고 있으며 또한 그 수행의 정도를 평가하는 다수의 연구가 실시된 바 있다(Bock, 1972; Douglas & Cohen, 2001; Glas & Suarez-Falcon, 2003; Kang & Chen, 2008; Liang & Wells, 2007; Orlando & Thissen, 2000, 2003; Sinharay, 2003, 2005; Stone, 2000; Stone & Zhang, 2003; Suarez-Falcon & Glas, 2003; Wells, 2004; Yen, 1981).

최근 들어 문항반응이론을 통한 대규모 검사 자료 분석에서 문항 수준에서의 모형 접합도 검정을 위하여 $S\text{-}X^2$ 지수를 사용하는 경우를 자주 볼 수 있다. 예를 들어, 한국교육종단연구 자료를 다양한 측면에서 분석한 김양분 외(2016)의 연구를 보면, 수직척도 개발을 위하여 2모수 문항반응모형을 선택하여 BILOG-MG((Zimowski, Muraki, Mislevy, & Bock, 1996) 프로그램을 사용하여 분석하였다. 하지만, 문항 적합도 검정 결과는 이 프로그램에서 제공하는 지수를 사용하지 않고 다른 방법으로 계산한 $S\text{-}X^2$ 지수를 보고하고 있음을 확인할 수 있다. Orlando와 Thissen(2000, 2003) 그리고 Kang과 Chen(2008)의 연구를 통하여 이 지수의 장점이 상당 부분 증명된 바 있지만, 이들 연구에서 고려한 표본크기의 제약으로 인하여 이와 같은 대규모 검사 자료를 분석할 때 $S\text{-}X^2$ 지수의 특성을 보다 명확히 확인할 필요가 있다. 따라서 본 연구에서는 IRT 분야에서 가장 신뢰할 수 있는 문항 적합도 검정 방법 중의 하나로서 널리 인정받고 있는 $S\text{-}X^2$ 지수에 대하여 주로 표본크기에 따른 영향 측면에서 보다 심층적으로 살펴보고자 한다.

$S\text{-}X^2$ 지수는 영가설 즉 '사용 모형이 주어진 문항 자료를 잘 설명한다'는 전제 하에서 카이제곱 분포를 따른다고 보고 통계적 유의성 검정 결과를 제공하는 형태로 활용된다. 그런데 교육통계 및 계량심리 분야에서 널리 알려져 있다시피, 카이제곱 분포에 기반한 통계적 검정 방법은 대개 표본크기 증가에 따라서 제1종 오류 가능성이 확대되는 경향이 있다. 이는 사례 수가 매우 많을 경우, 주어진 자료를 통계적 모형이 성공적으로 설명하고 있을지라

도 카이제곱 검증 결과가 모형-자료 간의 부적합을 가리키는 경우가 흔히 발생한다는 의미이다(Bentler and Bonnet, 1980; Jöoreskog and Söorbom, 1993). 따라서 Nisen과 Schwertman(2008)이 말하듯이 특정 목적을 위하여 카이제곱 검정을 실시하고자 할 때 적절한 수준의 표본크기를 파악하는 것이 중요한 이슈가 된다. 그러나 Orlando와 Thissen(2003) 그리고 Kang과 Chen(2008) 등의 연구에 따르면 IRT 적용 맥락에서 $S\text{-}X^2$ 지수는 기본적으로 카이제곱 통계치의 일종임에도 불구하고 표집 규모가 꽤 크다고 할 수 있는 n=5,000에 달할 때조차 제1종 오류의 발생 확률이 명목 유의수준에 매우 가깝게 유지됨을 보여주고 있다. 전자의 경우 고려된 사례 수는 500, 1000, 그리고 2000이었고, 후자의 경우 이에 더하여 5000을 더 고려하였다. 이 연구에서는 이러한 사실에 주목하여, 보다 다양한 수준의 표본크기에 따른 $S\text{-}X^2$ 문항 적합도 분석 방법의 수행 정도를 평가해 보고자 하였다.

검사 자료에 대한 IRT 분석은 그 특성상 흔히 대규모 검사 결과에 적용되기 때문에 사용 모형의 적합도를 살펴보고자 할 때 주어진 응답 자료 모두를 활용할지 아니면 일정 비율을 표집하여 사용할지를 결정할 필요가 있다. 따라서 표본크기가 $S\text{-}X^2$와 같은 모형 적합도 지수의 수행에 미치는 영향을 체계적으로 살펴보고 그 결과를 정리 및 확보하는 것은 검사 자료 분석가에게 큰 이점을 제공할 수 있을 것이다. 정리하여 다시 말하자면, 본 연구의 주된 목적은 주어진 검사 자료에 대한 다양한 이분 문항반응모형(dichotomous item response model) 적용 하에서 $S\text{-}X^2$ 지수를 통하여 문항 적합도를 검정하고자 할 때 표본크기가 이러한 통계적 유의성 검정의 제1종 오류에 미치는 영향을 살펴보는 데에 있다.

Ⅱ. 이론적 배경

1. 기존의 카이제곱 기반 문항 적합도 지수: BILOG-MG의 경우

BILOG-MG((Zimowski, Muraki, Mislevy, & Bock, 1996)는 문항 적합도 검정을 위하여 각 문항별로 카이제곱(χ^2) 통계치 값을 제공한다. 이 때 함께 계산된 자유도에 따라서 영가설(Ho: 문항 자료가 해당 모형에 의해서 잘 설명됨) 하에서의 유의확률이 제공된다. 이러한 유의확률 값이 연구자가 상정한 명목 유의수준보다 작을 경우 통계적 유의미성을 발견한 것이며 따라서 앞의 영가설을 부정하게 된다. BILOG-MG 프로그램의 매뉴얼에 따르면, 이 때 사용되는 명령어는 CHI(a, b) 형태가 되는데 여기서 a는 카이제곱 통계치 계산이 요구되는 문항수를 의미하고 b는 theta 능력 척도 상에서 등간격으로 선정되는 능력 값의 개수(구분구적점의 개수라고도 볼 수 있음)를 가리킨다. 그러나 실제 자료 분석을 위하여 BILOG-MG 3.0 프

로그램에서 이 명령어를 사용해 보면, a 값의 경우 전체 검사 문항 수 이하의 숫자를 지정하면 모든 문항 각각에 대한 카이제곱 통계치 및 검정 결과가 제공되며 전체 검사 문항 수보다 많은 값을 기입할 경우 카이제곱 통계치가 전혀 제공되지 않음을 확인할 수 있다. 또한 매뉴얼에서는 전체 검사 문항 수가 20개 미만일 경우 BILOG-MG 프로그램이 제공하는 카이제곱 통계치는 신뢰하기 어렵다고 밝히고 있다. b의 경우 디폴트 값이 9로 알려져 있으며 전체 문항 수가 20개보다 적을 경우 9보다 큰 값을 기입하여도 9가 적용된다.

BILOG-MG 프로그램에서 제공하는 카이제곱 통계치는 Yen(1981)이 제시한 식 (1)과 같은 Q_1 통계치와 기본적으로 같다고 볼 수 있으며 이하에서는 이를 Q_1-χ^2로 부르기로 한다. 다만 Q_1 의 경우 자유도 계산 시 위의 b값에서 문항모수의 수만큼 감하는 반면 BILOG-MG 카이제곱 통계치 Q_1-χ^2는 이러한 고려를 하지 않는다는 차이가 있다. 또한 특정 문항 i를 위한 카이제곱 통계치를 계산하기 위한 이 식에서 K는 BILOG-MG 프로그램의 CHI=(a,b)에서 b와 같다고 볼 수 있다. Yen은 Q_1 통계치 계산을 위하여 K=10을 사용한 바 있다. 이 식이 함의하는 바는, 첫째, '능력모수 추정치(theta parameter estimates)'를 바탕으로 피험자들을 K 개의 동질적 집단들로 나눈다는 것이며, 둘째, 각 집단에서 관찰된 정답 비율($= O_{k1} = 1 - O_{k0}$)과 자료 분석을 위하여 사용된 문항반응모형에 의해서 계산된 혹은 기대된 정답 비율($= E_{k1} = 1 - E_{k0}$)을 이용하여 카이제곱 통계치를 계산한다는 것이다. N_k는 각 집단에 속한 피험자들의 빈도를 의미한다.

$$Q_1 = \sum_{k=1}^{K} \sum_{z=0}^{1} N_k \frac{(O_{ikz} - E_{ikz})^2}{E_{ikz}} = \sum_{k=1}^{K} N_k \frac{(O_{ikz} - E_{ikz})^2}{E_{ikz}(1 - E_{ikz})} \quad (1)$$

2. S-X^2 문항 적합도 지수 및 그 속성

하나의 문항 i를 위한 문항 적합도 지수 S-X^2를 구하기 위한 계산 공식은 아래 식 (2)와 같다. 이 식에서 F는 각 검사 문항의 범주가 $z = 0, 1, \ldots Z_i$때 검사의 최고 점수 즉 $\sum Z_i$를 뜻한다. 모든 문항에 대하여 오답은 0 그리고 정답은 1로 채점되는 상황 하에서 F 혹은 $\sum Z_i$의 값은 총 문항 수 I가 된다. 이 때 검사 점수는 0점부터 I점 만점까지 존재하게 되는데, 기존 BILOG-MG 프로그램에서 제공되는 카이제곱 기반의 문항 적합도 지수(Q_1-χ^2)와 달리, S-X^2는 이러한 실제 검사 점수 각각을 집단 구분을 위한 수단으로 사용한다. 카이제곱 값의 계산은 보통 관찰값과 기댓값의 비교를 통하여 이루어지는데, 식 (1)과 같은 방식으로 계산되는 문항 적합도 지수는 능력모수 추정치를 기반으로 피험자들을 몇 개의 집단으로

나눈 뒤 관찰 비율 혹은 빈도를 구함에 따라 순수한 관찰값을 이용하여 계산되었다고 보기 어렵다는 문제를 가지고 있었다(Orlando & Thissen, 2000).

검사 점수에 따른 각 집단을 k=0, 1, 2, ..., I 라고 볼 수 있을 때, S-X^2의 계산에서 0과 I 집단은 제외된다. 왜냐하면 검사점수가 0점인 집단에 속하는 피험자들은 모든 문항을 틀렸다는 것이기 때문에 그 중에서 해당 문항을 맞힌 사람이 있을 수 없을 것이며, 반대로 검사점수가 I점인 집단에 속하는 피험자들은 모든 문항을 맞혔기 때문에 해당 문항을 다 맞혔을 것이기 때문이다.

$$S-X^2=\sum_{k=Z_i}^{F-Z_i}\sum_{z_i=0}^{Z_i}N_k\frac{(O_{ikz}-E_{ikz})^2}{E_{ikz}}=\sum_{k=1}^{F-1}\sum_{z_i=0}^{1}N_k\frac{(O_{ikz}-E_{ikz})^2}{E_{ikz}} \quad (2)$$

이 식에서 N_k는 k 점수 집단에 속하는 피험자 수를 의미한다. O_{ikz}는 문항 i에서 k 검사 점수 집단에서 문항범주 z(이분 문항에서의 문항점수 0 혹은 1)에 속하는 관찰 비율을 의미하며, E_{ikz}는 같은 경우에 적합도를 검증하려는 문항반응모형을 통해 계산된 기대 비율을 의미한다. 즉 같은 검사 점수를 받은 피험자들을 동일 집단에 속한 것으로 보고 각 문항 범주마다의 관찰된 반응 비율과 문항반응모형에 의해서 계산된 기대 비율을 사용하여 한 문항의 모형 적합에 관련된 카이제곱 값을 계산하는 것이다. 이 때 대응되는 자유도는 일차적으로 유관표의 전체 셀 수에서 문항모수의 수와 1보다 작은 빈도를 갖는 셀의 수를 감하여 얻게 되며, 다시 문항모수의 수만큼 감한 값이 최종 자유도로 사용된다. 위의 식에서 E_{ikz}를 구체적으로 어떻게 계산하는지 등을 포함하여 S-X^2의 계산 방법에 대한 보다 자세한 내용은 Orlando와 Thissen(2000)에 기술되어 있다.

S-X^2 지수는 기본적으로 카이제곱 검증을 통하여 문항의 적합도에 대한 영가설(Ho: 문항 자료가 해당 모형에 의해서 잘 적합된다)을 검정하기 때문에, 명목 유의수준에 해당하는 카이제곱 값 다시 말하여 임계값(critical value)에 비하여 그 값이 지나치게 클 때 영가설을 부정하게 된다. 또한 모의실험을 통하여 밝혀진 바에 따르면 경험적 제1종 오류 및 통계적 검정력에 있어서 다른 문항적합도 지수에 비해 매우 우수한 결과를 산출하는 것으로 알려져 있다(Orlando & Thissen, 2000, 2003; Kang & Chen, 2008). 그럼에도 불구하고, S-X^2 지수 역시 카이제곱 검증에 기반을 하기 때문에 발생하는 문제점으로서 표본크기 혹은 사례 수가 커질수록 제1종 오류가 점차적으로 증가하는 문제점이 언젠가는 나타날 것으로 추측해 볼 수 있다. 하지만 대규모 학업성취도 검사 자료 분석 등에 있어서 표본크기가 매우 클 경우 S-X^2 지수가 제대로 기능하는지에 대한 체계적 연구가 충분히 이루어지지 않은 것으로 보인다. 따라서 표본크기를 다양하게 고려하여(말하자면, 본 연구에서는 200부터 100,000까지 포함)

제1종 오류가 제대로 통제되는 사례 수 범위가 어디까지인지를 탐색해 볼 필요가 있다.

Ⅲ. 연구 방법

1. 모의실험 설계 및 자료 생성

표본크기가 $S\text{-}X^2$ 지수의 수행 즉 경험적 제1종 오류 정도나 통계적 검정력 등에 미치는 영향을 살펴보기 위하여, 본 연구에서는 다음과 같은 모의실험 연구를 실시하였다. 첫째, 이분 문항 검사 자료를 생성하기 위한 문항반응모형으로 1모수-, 2모수-, 3모수 로지스틱 모형(이하에서는 각각 1PLM, 2PLM, 그리고 3PLM으로 부르기로 함)을 고려하였다. 둘째, 다양한 교육 및 심리 검사 상황 하에서 하나의 검사가 포함하는 문항 수가 다양할 수 있음을 감안하여 총 문항 수가 10개, 30개, 그리고 50개인 상황을 상정하였다. 셋째, 기본적 연구 목적에 충실하고자 표본크기(n) 혹은 피험자 수를 200, 500, 1000, 2000, 4000, 6000, 8000, 10000, 20000, 40000, 60000, 80000, 그리고 100000과 같이 13개 경우를 고려하였다. 결과적으로 본 연구에서의 모의실험 조건 수는 총 117개(=3 종류의 자료 생성 문항반응모형 × 3 종류의 검사 문항 수 × 13 종류의 표본크기)이다.

<표 1>에서는 모의실험 자료 생성을 위하여 사용된 문항모수가 제시되어 있다. 이들은 2012년에 실시된 시도 수준의 고1 수학 학업성취도 검사 자료로부터 확보되었으며 문항변별도의 평균은 1 그리고 문항난이도의 평균은 0으로 조정하여 사용하였다. 2모수 모형으로 자료 생성 시에는 모든 문항추측도를 0으로 보았으며, 다시 1모수 모형으로 자료 생성 시에는 이에 더하여 모든 문항의 변별도가 1이 되도록 하였다. 30문항 및 10문항 조건의 자료를 생성할 때에는 이들 50문항 중에서 무선적으로 각각 30문항과 10문항을 선택한 뒤 다시 문항변별도의 평균은 1 그리고 문항난이도의 평균은 0이 되도록 조정하여 사용하였다.

117개 각각의 모의실험 조건 당 100개의 반복 자료(replicated data sets)가 생성되었다. 자료생성은 IRT를 사용한 모의실험 연구에서 흔히 사용되는 다음과 같은 3 단계 절차를 통하여 이루어졌다. 첫째, <표 1>에서 제시된 진문항모수와 $N(0,1)$으로부터 무선 표집된 진능력모수를 이용하여 각 문항에 대하여 개별 가상 피험자가 정답 반응할 확률을 계산하였다. 둘째, 이러한 값을 균일분포 $U(0,1)$에서 무선적으로 생성된 값과 비교하여 더 클 경우는 맞는 것으로 그리고 더 작을 경우는 틀린 것으로 기록하였다. 셋째, 하나의 모의실험 조건 하에서 반복자료를 생성할 때에는 동일한 진문항모수들을 사용하였으나 피험자 능력모수는 매번 $N(0,1)$으로부터 무선적으로 새로 추출하였다. 이를 통하여 마치 동일한 검사를 동일 능력분

〈표 1〉 자료 생성을 위한 실제 검사 문항의 모수

문항	a	b	c	문항	a	b	c
1	0.999	-0.323	0.134	26	0.809	-2.104	0.117
2	1.182	-0.196	0.117	27	0.545	-1.528	0.171
3	0.644	-0.001	0.131	28	1.103	0.903	0.209
4	0.544	-0.528	0.095	29	0.805	0.005	0.161
5	0.639	-2.132	0.110	30	0.994	0.923	0.085
6	1.236	-0.025	0.334	31	1.162	1.003	0.184
7	1.323	1.037	0.078	32	1.323	0.458	0.270
8	1.329	0.567	0.084	33	0.525	-0.938	0.147
9	0.693	-0.557	0.067	34	0.847	-0.296	0.136
10	1.218	1.480	0.176	35	0.767	-1.634	0.163
11	0.501	1.123	0.130	36	0.956	-0.453	0.293
12	1.092	-0.630	0.296	37	0.607	0.058	0.385
13	1.081	0.236	0.386	38	1.213	-1.280	0.145
14	1.090	0.083	0.163	39	1.303	0.339	0.228
15	1.567	0.550	0.104	40	0.651	0.168	0.086
16	1.105	-1.049	0.266	41	0.893	-0.376	0.167
17	0.715	1.627	0.233	42	1.634	1.258	0.175
18	0.686	-1.588	0.127	43	1.028	1.355	0.165
19	1.555	0.140	0.078	44	0.951	-0.215	0.089
20	0.650	-1.813	0.173	45	0.940	1.111	0.130
21	0.690	0.909	0.249	46	1.020	-0.792	0.180
22	1.062	-0.498	0.286	47	0.854	-0.467	0.105
23	1.442	0.953	0.081	48	1.684	1.492	0.213
24	1.527	1.775	0.177	49	1.367	0.186	0.223
25	0.657	0.110	0.254	50	0.929	-0.420	0.118

포를 가진 100개의 다른 피험자 집단이 치룬 것과 같은 모의실험 자료를 각 모의실험 조건에서 생성할 수 있었다.

2. 문항 적합도 지수 계산 및 추정모형 사용 계획

생성된 각 모의실험 자료에 대한 문항모수 추정과 식 (1)에 따른 Q_1-χ^2 즉 IRT 분야에서의 전통적 카이제곱 통계치 계산 및 문항 적합도 결과 도출은 BILOG-MG 프로그램을 통하여 이루어졌다. 이를 위한 BILOG-MG 예시 코드는 [부록]에 제시된 바와 같다. 여기에서는 50문항 조건들에서 사용된 1모수, 2모수, 3모수 문항반응모형 각각을 위한 코드들을 제시하고 있다. 각 코드에서 발견할 수 있는 바와 같이 CHI=(a,b)의 명령어에서 a에 대해서는 검사 문항수를 기입하였고 b에 대해서는 항상 디폴트 값인 9를 사용하였다. 앞에서 제시된 식 (2)에 따른 S-X^2 지수는 연구자에 의하여 MATLAB으로 작성된 프로그램을 통하여 계산된다. 관련 코드는 부록으로 제시되기에는 그 양이 많기 때문에 독자의 요청이 있을 경우 이메일을 통하여 제공될 수 있다. 또한 모의실험 연구를 효율적으로 수행하기 위하여 본 연구에서 사용하지는 않았으나, 근래 들어서 많이 사용되고 있는 프로그램 중 하나인 IRTPRO(Cai, Thissen, du Toit, 2011)를 통하여 S-X^2 지수의 계산이 가능한 것으로 알려져 있다.

각 모의실험 조건에서 생성모형(generating model, GM)이 무엇인가에 따라서 문항모수 추정 및 문항 적합도 지수 계산을 위한 추정모형(calibrating model, CM)이 <표 2>와 같이 사용되었다. 이 그림은 Orlando와 Thissen(2000)이 제1종 오류와 통계적 검정력 결과를 도출하기 위하여 사용한 모형 사용 계획과 동일하다. 다시 말하여, GM과 CM이 동일할 때 어떤 문항의 모형 적합도 검정 결과 영가설이 기각된다면 이를 제1종 오류로 볼 수 있다. 하지만 GM보다 더 단순한 모형을 CM으로 사용할 경우 모형 적합도에 문제가 생기는 것이 당연하다고 기대되기 때문에 이 경우 영가설이 기각된다면 이를 통계적 검정력이 좋은 것으로 판단할 수 있다. <표 2>에서 제시된 계획에 따라서 GM 보다 더 복잡한 모형을 CM으로 사용하지는 않는다.

〈표 2〉 생성모형에 따른 추정모형 사용 계획 (Orlando & Thissen, 2000, p.55)

추정모형 (CM)	생성모형 (GM)		
	1PLM	2PLM	3PLM
1PLM	제1종 오류 (GM=CM)	통계적 검정력 (GM>CM)	통계적 검정력 (GM>CM)
2PLM	-	제1종 오류 (GM=CM)	통계적 검정력 (GM>CM)
3PLM	-	-	제1종 오류 (GM=CM)

특정 모의실험 조건에서 문항 수가 10개인 경우 100개의 반복 자료 모두를 고려할 때 총 1,000개의 문항이 존재하며 이 중 몇 개의 문항이 $S\text{-}X^2$ 및 $Q_1\text{-}\chi^2$ 지수 계산 결과 통계적으로 유의미한 지 그 비율을 구한다. 본 연구에서는 두 적합도 지수 모두에 대하여 통계적 유의성 검정을 위하여 명목 유의수준 0.05를 사용하였다. 이렇게 각 조건에서 계산된 비율(=총 부적합 문항 수 / 총 문항 수)은 생성모형과 추정모형이 일치하는 경우에서는 제1종 오류 수준으로 이해할 수 있으며, 추정모형이 생성모형보다 단순한 경우 통계적 검정력으로 볼 수 있다. 특정 모의실험 조건의 문항 수가 30개와 50개인 경우, 한 조건에서 고려되는 총 문항 수는 각각 3,000개와 5,000개가 된다.

Ⅳ. 연구 결과

1. 경험적 제1종 오류의 수준

명목 유의수준 0.05 적용 하에서, 자료 생성모형과 추정모형이 일치할 때(GM=CM) 각 조건에서 얼마나 많은 문항들이 식 (1)의 $Q_1\text{-}\chi^2$과 식 (2)의 $S\text{-}X^2$ 지수에 의하여 부적합 혹은 통계적으로 유의미한 것으로 판정되는지의 비율을 구한 결과는 <표 3>, <표 4> 그리고 <표 5>에 제시된 바와 같다. 각 적합도 지수의 수행을 평가함에 있어서, 제1종 오류가 0.05에 가까울수록 바람직하다고 볼 수 있다. 이들 표에서는 다소 임의적일 수 있으나 공인된 기준이 존재하는 것은 아닌 바, 0.08보다 큰 경우를 과도한 팽창(inflation) 즉 제1종 오류 수준이 제대로 통제되고 있지 못한 것으로 보고 해당 배경 부분을 어둡게 처리하였다. 일부 선행연구에서는 명목 유의수준이 0.05인 경우 경험적 제1종 오류 수준이 적정한 수준에서 통제되고 있다고 볼 수 있는지를 다음과 같은 일방적 95% 신뢰구간(confidence interval, CI)을 사용하여 평가하기도 한다. 즉 경험적 제1종 오류가 0과 $0.05+1.645\sqrt{(0.05\times0.95)/(I\times R)}$ 사이에서 유지되는가를 보는 것이다(Stone & Hensen, 2000). 여기에서 R은 한 모의실험 조건에서 반복 생성된 검사자료의 수(본 연구에서는 100)를 의미하며 I는 한 모의실험 조건에서 검사가 갖는 문항 수를 의미한다. 이 때 각 조건의 문항 수인 10, 30, 혹은 50에 따라서 각각 [0, 0.061], [0, 0.057], 그리고 [0, 0.055]의 CI들을 기준으로 판단하는 것도 가능할 것이다. 하지만, 본 연구에서는 이와 같은 기준들이 너무 보수적일 수 있다고 판단하였다. 동시에 기존 Orlando와 Thissen(2000)의 연구에서는 0.07 정도까지 별다른 문제가 없는 것으로 보며 다소 큰 값인 0.09 혹은 0.10을 기대보다 약간 큰 정도의 결과로 해석하는 경향이 있는 것을 참고하여, 경험적 제1종 오류에 대한 적절성 판단을 위하여 0.08을 일률적으로 해석 기준으로 사

용하였다. 하지만 이러한 기준의 임의성으로 인하여 <표 3> 등을 해석함에 있어서 독자의 주의가 요구됨이 사실이다.

이들 표를 통하여 나타난 $S\text{-}X^2$ 지수의 수행을 살펴보면, 표본크기가 매우 작은 200명 수준인 경우를 제외하고 거의 모든 조건에서 제1종 오류가 적정 수준에서 통제되고 있음을 확인할 수 있다. 즉 문항반응모형을 통한 이분문항 검사 자료 분석 시 피험자 수가 500명을 넘는다면 적용 모형 및 문항 수에 관계없이 경험적 제1종 오류가 명목 유의수준에 가깝게 유지될 수 있음을 알 수 있다. $S\text{-}X^2$ 지수의 경우 <표 4>에서 확인할 수 있는 바와 같이 문항 수가 30개일 때는 모든 모의실험 조건에서 만족할 만한 수준으로 제1종 오류가 통제되는 것으로 나타났지만 <표 5>의 문항 50개 조건에서는 표본크기가 200일 때 1PLM과 3PLM에서 제1종 오류가 크게 나타나는 문제가 있었다. 이를 통하여 표본크기가 매우 작을 경우 문항반응모형 적용 및 $S\text{-}X^2$ 지수의 수행에 주의가 요구됨을 알 수 있다.

〈표 3〉 경험적 제1종 오류 수준: 생성모형과 추정모형이 일치하는 경우 (문항 수=10)

표본크기 (n)	$S\text{-}X^2$			$Q_1\text{-}\chi^2$		
	1PLM	2PLM	3PLM	1PLM	2PLM	3PLM
200	0.038	0.050	0.104	0.036	0.037	0.136
500	0.058	0.062	0.073	0.166	0.220	0.412
1,000	0.057	0.038	0.072	0.505	0.632	0.701
2,000	0.056	0.049	0.074	0.938	0.931	0.843
4,000	0.047	0.060	0.071	0.999	0.997	0.927
6,000	0.044	0.045	0.062	1.000	1.000	0.967
8,000	0.042	0.049	0.069	1.000	1.000	0.970
10,000	0.044	0.058	0.062	1.000	1.000	0.975
20,000	0.053	0.056	0.070	1.000	1.000	0.999
40,000	0.052	0.059	0.058	1.000	1.000	1.000
60,000	0.052	0.044	0.045	1.000	1.000	1.000
80,000	0.054	0.050	0.058	1.000	1.000	1.000
100,000	0.046	0.054	0.056	1.000	1.000	1.000

검사 문항 수가 10개인 경우에 대하여 조사된 <표 3>을 보면 BILOG-MG 프로그램을 통하여 계산된 카이제곱 적합도 지수 $Q_1\text{-}\chi^2$의 경우 제1종 오류의 부적절한 팽창이 거의 모든 표본크기에서 발생하고 있음을 알 수 있다. 이는 BILOG-MG 프로그램의 매뉴얼에서 언급하

〈표 4〉 경험적 제1종 오류 수준: 생성모형과 추정모형이 일치하는 경우 (문항 수=30)

표본크기	$S\text{-}X^2$			$Q_1\text{-}\chi^2$		
(n)	1PLM	2PLM	3PLM	1PLM	2PLM	3PLM
200	0.058	0.059	0.075	0.018	0.018	0.016
500	0.061	0.050	0.055	0.028	0.019	0.026
1,000	0.054	0.054	0.063	0.046	0.029	0.070
2,000	0.051	0.052	0.060	0.078	0.068	0.174
4,000	0.052	0.053	0.060	0.173	0.199	0.372
6,000	0.056	0.053	0.062	0.277	0.322	0.472
8,000	0.050	0.055	0.060	0.401	0.456	0.526
10,000	0.056	0.047	0.064	0.490	0.545	0.583
20,000	0.047	0.053	0.056	0.802	0.777	0.705
40,000	0.053	0.049	0.056	0.973	0.954	0.836
60,000	0.047	0.049	0.050	0.997	0.996	0.910
80,000	0.051	0.058	0.056	1.000	1.000	0.953
100,000	0.058	0.049	0.058	1.000	1.000	0.972

〈표 5〉 경험적 제1종 오류 수준: 생성모형과 추정모형이 일치하는 경우 (문항 수=50)

표본크기	$S\text{-}X^2$			$Q_1\text{-}\chi^2$		
(n)	1PLM	2PLM	3PLM	1PLM	2PLM	3PLM
200	0.090	0.072	0.086	0.017	0.014	0.009
500	0.061	0.053	0.064	0.020	0.015	0.018
1,000	0.059	0.058	0.070	0.031	0.018	0.025
2,000	0.056	0.051	0.055	0.041	0.024	0.039
4,000	0.059	0.049	0.059	0.052	0.040	0.089
6,000	0.054	0.060	0.055	0.076	0.063	0.137
8,000	0.058	0.056	0.057	0.101	0.094	0.204
10,000	0.056	0.049	0.058	0.119	0.120	0.276
20,000	0.052	0.046	0.052	0.252	0.337	0.494
40,000	0.055	0.050	0.055	0.552	0.641	0.639
60,000	0.048	0.045	0.057	0.662	0.775	0.711
80,000	0.050	0.053	0.063	0.850	0.844	0.768
100,000	0.048	0.046	0.059	0.911	0.884	0.805

고 있는 바와 같이, 전체 문항 수가 20개보다 적을 경우 이러한 적합도 지수를 신뢰하기 어렵다는 점을 단적으로 보여주고 있는 것으로 보인다. 검사 문항 수가 30개인 경우, <표 4>에서 볼 수 있는 바와 같이 표본크기가 1,000명인 경우 Q_{I}-χ^2이 적정 수준에서 제1종 오류를 통제하고 있음을 알 수 있으며 그보다 표본크기가 커질 경우 3PLM에 대한 문항 적합도 검정이 제대로 이루어지기 어려움을 보여주고 있다. 검사 문항 수가 50개인 경우에 대한 <표 5>의 결과를 보면, 표본크기가 2,000명인 경우까지 Q_{I}-χ^2이 적정 수준에서 제1종 오류를 통제하고 있음을 알 수 있으며 그보다 표본크기가 커질 경우 1PLM과 2PLM 적용 하에서는 6,000명인까지 어느 정도 통제가 이루어졌으나 3PLM 하에서는 문항 적합도 검정이 제대로 이루어지지 않음을 알 수 있었다.

2. 통계적 검정력 수준

앞의 <표 2>에서 언급한 바와 같이 생성모형이 추정모형보다 더 복잡한 혹은 더 많은 문항모수를 사용하는 경우(GM>CM), 본 연구에서는 문항 적합이 제대로 이루어지지 않는

〈표 6〉 통계적 검정력 수준: 생성모형이 추정모형 보다 복잡한 경우 (문항 수=10)

표본크기 (n)	S-X^2			Q_{I}-χ^2		
	GM > CM					
	2PLM > 1PLM	3PLM > 1PLM	3PLM > 2PLM	2PLM > 1PLM	3PLM > 1PLM	3PLM > 2PLM
200	0.147	0.140	0.081	0.149	0.150	0.047
500	0.312	0.225	0.091			
1,000	0.530	0.395	0.131			
2,000	**0.774**	0.579	0.233			
4,000	**0.904**	**0.765**	0.394			
6,000	**0.964**	**0.807**	0.502			
8,000	**0.989**	**0.831**	0.603			
10,000	**0.999**	**0.863**	0.661			
20,000	**1.000**	**0.925**	**0.843**			
40,000	**1.000**	**0.980**	**0.974**			
60,000	**1.000**	**0.992**	**0.998**			
80,000	**1.000**	**1.000**	**1.000**			
100,000	**1.000**	**1.000**	**1.000**			

것이 당연하다고 보고 식 (1)의 Q_1-χ^2과 식 (2)의 S-X^2 지수에 의하여 부적합 문항으로 판정되는 비율을 통계적 검정력 수준으로 보았다(Orlando & Thissen, 2000). 통계적 검정력 관련 연구 결과는 <표 6>, <표 7>, 그리고 <표 8>에 제시되어 있으며, 앞의 <표 3>, <표 4>, 그리고 <표 5>의 연구 결과를 통하여 제1종 오류가 제대로 통제되지 않는 조건에 대해서는 해석상의 불필요한 혼동을 막기 위하여 통계적 검정력 수준에 대한 결과를 보고하지 않기로 하였다. 또한 <표 6>, <표 7>, 그리고 <표 8>에서는 통계적 검정력 수준이 0.7을 넘는 경우에 대하여 굵은 글씨로 표시하였다. 통계적 검정력은 100%에 가까울수록 바람직하지만 경험적 연구에서 이러한 결과를 얻기는 매우 어려우며, 연구에 따라서 80%를 바람직하다고 보거나(e.g., 이효진, 김양수, 박인, 2013) 적절한 수준으로 70%를 언급하는 연구도 존재한다(e.g., Sham & Purcell, 2014). 본 연구에서는 후자의 기준에 따르기로 하였다.

검사 문항 수가 10개인 경우에 대한 <표 6>에 따르면, Q_1-χ^2의 경우 제1종 오류가 어느 정도 통제되는 n=200에서조차 통계적 검정력이 매우 낮은 것으로 나타났기 때문에 적어도 문항 수가 10개인 경우 문항 적합도 검정을 위하여 고려되기 어려운 적합도 지수인 것을 확

〈표 7〉 통계적 검정력 수준: 생성모형이 추정모형 보다 복잡한 경우 (문항 수=30)

표본크기 (n)	S-X^2			Q_1-χ^2		
	GM > CM					
	2PLM > 1PLM	3PLM > 1PLM	3PLM > 2PLM	2PLM > 1PLM	3PLM > 1PLM	3PLM > 2PLM
200	0.172	0.159	0.074	0.119	0.143	0.026
500	0.269	0.285	0.075	0.365	0.375	0.041
1,000	0.425	0.430	0.126	0.595	0.582	0.125
2,000	0.642	0.619	0.210	**0.803**	**0.727**	0.332
4,000	**0.827**	**0.782**	0.353			
6,000	**0.895**	**0.828**	0.454			
8,000	**0.928**	**0.864**	0.534			
10,000	**0.938**	**0.877**	0.602			
20,000	**0.960**	**0.946**	**0.803**			
40,000	**0.987**	**0.971**	**0.935**			
60,000	**0.998**	**0.973**	**0.970**			
80,000	**1.000**	**0.975**	**0.983**			
100,000	**1.000**	**0.977**	**0.988**			

인할 수 있었다. $S\text{-}X^2$ 지수의 경우 1PLM을 추정모형으로 사용할 때 부적합의 정도에 따라서 표본크기가 2,000 혹은 4,000 이상일 때 충분한 수준의 통계적 검정력을 가질 수 있는 것으로 보였다. 2PLM 적용 시에는 표본크기가 20,000명 이상일 때 84.3% 이상의 통계적 검정력을 갖는 것으로 나타났다. <표 7>에서는 검사 문항이 30개인 경우, 1PLM을 추정모형으로 사용할 때 $Q_1\text{-}\chi^2$이 충분한 통계적 검정력을 가지려면 피험자 수가 2,000명 정도가 되어야 하는 것으로 나타났다. $S\text{-}X^2$ 지수의 경우 1PLM을 추정모형으로 사용할 때 4,000명 이상 그리고 2PLM을 추정모형으로 사용할 때 20,000명 이상의 피험자가 있어야 일정 수준 이상의 통계적 검정력을 확보할 수 있는 것으로 나타났다. 검사 문항이 50개인 경우에 대한 <표 8>을 보면, $Q_1\text{-}\chi^2$은 70% 이상의 통계적 검정력 확보를 위하여 1PLM의 경우 2,000~4,000명 정도가 필요한 것으로 나타났으며 2PLM은 6,000명 정도가 필요한 것으로 보였다. $S\text{-}X^2$ 지수의 경우 검사 문항 수가 30인 경우와 50인 경우 그 결과가 매우 유사하였다.

〈표 8〉 통계적 검정력 수준: 생성모형이 추정모형 보다 복잡한 경우 (문항 수=50)

표본크기 (n)	$S\text{-}X^2$			$Q_1\text{-}\chi^2$		
	GM > CM					
	2PLM > 1PLM	3PLM > 1PLM	3PLM > 2PLM	2PLM > 1PLM	3PLM > 1PLM	3PLM > 2PLM
200			0.085	0.113	0.128	0.018
500	0.232	0.260	0.089	0.340	0.340	0.052
1,000	0.346	0.403	0.134	0.549	0.567	0.135
2,000	0.521	0.586	0.220	0.697	**0.761**	0.283
4,000	0.672	**0.780**	0.355	**0.834**	**0.883**	0.572
6,000	**0.752**	**0.853**	0.462	**0.887**	**0.935**	**0.745**
8,000	**0.797**	**0.892**	0.567			
10,000	**0.832**	**0.917**	0.643			
20,000	**0.902**	**0.971**	**0.829**			
40,000	**0.961**	**0.982**	**0.955**			
60,000	**0.988**	**0.984**	**0.988**			
80,000	**0.996**	**0.985**	**0.995**			
100,000	**0.999**	**0.986**	**0.998**			

Ⅴ. 결론 및 논의

카이제곱 분포를 따르는 $S\text{-}X^2$ 문항 적합도 지수가 제1종 오류 가능성에 있어서 표본크기의 영향을 어느 정도 받는지를 모의실험 연구를 통해 살펴보는 것이 본 연구의 주된 목적이었다. 이에 추가적으로, 본 연구에서는 전통적으로 활용되어 온 문항 적합도 지수인 $Q_1\text{-}\chi^2$의 수행도 함께 조사하였으며 다양한 이분 문항반응모형 및 문항 수도 함께 고려하여 문항 적합도에 대한 통계적 유의성 검정이 어느 경우에 제대로 이루어질 수 있는지를 확인하고자 하였다. 최근 들어 $S\text{-}X^2$ 지수는 다양한 대규모 검사 자료 분석을 위하여 활발하게 사용되고 있다. 예를 들어, 앞에서 언급한 한국교육종단연구 자료를 대상으로 한 수직척도 개발 연구(김양분 외, 2016) 외에도 미국 미시간 주 교육부(Michigan Department of Education, 2011)의 학업성취도 검사(Michigan Merit Exam, MME) 자료 분석 등에서 $S\text{-}X^2$ 지수가 사용되고 있다. 하지만 많은 피험자를 포함한 자료를 다루는 경우 실제 그 수행이 명확히 밝혀진 바가 없기 때문에 본 연구에서는 이를 해결하고자 하였다.

연구 결과에 따르면, $S\text{-}X^2$ 지수는 표본크기가 100,000명까지 극단적으로 커질 때조차 제1종 오류의 과도한 팽창 없이 명목 유의수준인 5% 크기 근처에서 제1종 오류를 통제할 수 있는 것으로 나타났다. 통계적 검정력 역시 2,000 내지는 4,000명 이상일 때 70% 이상으로 나타나는 경우를 자주 볼 수 있었다. 하지만 BILOG-MG 프로그램의 문항 적합도 지수는 검사 문항 수가 10개 정도로 작을 때에는 사용해서는 안 되는 것으로 나타났으며, 문항반응모형에 따라서 약간의 차이가 있으나 대체로 문항 수가 30개나 50개로 충분할 때 2,000~6,000명 정도의 표본크기에서만 제1종 오류 정도 및 통계적 검정력이 만족스럽게 보이는 몇몇 경우가 있었다(<표 6>, <표 7>, <표 8> 참조).

그런데 본 연구에서 제시하는 BILOG-MG 프로그램에 의한 문항 적합도 검정 결과를 해석함에 있어서 다음과 같은 사항을 주의할 필요가 있다. 앞에서 밝힌 바와 같이 그리고 [부록]에서 확인할 수 있는 바와 같이, 본 연구에서는 CHI=(a, b)의 명령어에서 a에 대해서는 검사 문항 수를 그리고 b에 대해서는 항상 디폴트 값인 9를 사용하였다. 이는 피험자 수가 아무리 많더라도 추정된 피험자 능력 모수를 바탕으로 언제나 피험자들을 9개의 동질집단으로 나누어서 카이제곱 검정을 실시한다는 뜻이 된다. 이러한 집단 구분은 카이제곱 분포의 자유도에 영향을 미치게 되며, 더 나아가 통계적 유의성 검정 결과에도 영향을 미치게 된다.

<표 9>에서는 50문항 검사에 대하여 100,000명이 응답한 한 모의실험 자료에 대하여 생성모형과 추정모형이 모두 3PLM인 경우(GM=CM=3PLM)의 문항 적합도 검정 결과를 제시하고 있다. 이 표의 가장 오른쪽 열을 보면 $S\text{-}X^2$ 지수에 따른 결과를 확인할 수 있는데 예상할 수 있는 바와 같이 거의 문항들(문항 1, 11, 14만 부적합으로 나타남: 제1종 오류)이

〈표 9〉 GM=3PLM, 50문항, 100000명 조건에서 3PLM 문항 적합도 검정 결과

item	Q_{I}-χ^2 (BILOG-MG가 제공하는 문항 적합도 지수)									S-X^2		
	통계치 (b=9)	자유도	유의확률	통계치 (b=20)	자유도	유의확률	통계치 (b=40)	자유도	유의확률	통계치	자유도	유의확률
1	95.4	9	0.000	136.1	20	0.000	151.7	40	0.000	58.735	42	0.045
2	103.4	9	0.000	142.6	20	0.000	156.8	40	0.000	26.752	42	0.968
3	56.5	9	0.000	77.4	20	0.000	85.5	40	0.000	31.474	42	0.882
4	93.7	9	0.000	103.1	20	0.000	114.1	40	0.000	41.834	42	0.478
5	165.5	9	0.000	182.8	20	0.000	191.3	40	0.000	35.185	43	0.796
6	46.5	9	0.000	80.5	20	0.000	95.7	40	0.000	50.753	42	0.167
7	14.8	9	0.097	30.2	20	0.066	43.7	40	0.316	36.060	41	0.690
8	30.4	9	0.000	50.5	20	0.000	63.7	40	0.010	41.895	41	0.432
9	105.3	9	0.000	130.8	20	0.000	144.6	40	0.000	32.381	42	0.857
10	5.7	9	0.766	16.7	20	0.670	50.7	40	0.119	37.706	42	0.660
11	16.6	9	0.055	38.4	20	0.008	55	40	0.057	58.376	42	0.048
12	123.8	9	0.000	142.9	20	0.000	156.2	40	0.000	46.261	41	0.264
13	20.6	9	0.015	42.2	20	0.003	55.2	40	0.055	42.426	42	0.453
14	48.1	9	0.000	62.6	20	0.000	84.8	40	0.000	58.798	42	0.044
15	49.5	9	0.000	93	20	0.000	109	40	0.000	46.487	41	0.257
16	198.7	9	0.000	220.9	20	0.000	259.3	40	0.000	44.534	41	0.325
17	6.8	9	0.653	18.8	20	0.535	35.2	40	0.687	43.575	42	0.404
18	183.5	9	0.000	200.3	20	0.000	211.1	40	0.000	43.594	42	0.403
19	119.7	9	0.000	158	20	0.000	181.9	40	0.000	43.796	40	0.314
20	166.9	9	0.000	181	20	0.000	202	40	0.000	44.412	43	0.412
21	10	9	0.351	25.8	20	0.171	42.9	40	0.347	39.801	42	0.568
22	82.3	9	0.000	115	20	0.000	124.5	40	0.000	47.750	42	0.251
23	20.8	9	0.014	45.1	20	0.001	65	40	0.008	33.912	41	0.776
24	17.1	9	0.047	24.1	20	0.239	44.4	40	0.290	35.056	42	0.767
25	22.5	9	0.007	38.7	20	0.007	57	40	0.039	28.508	42	0.944
26	243.6	9	0.000	268.9	20	0.000	267.1	40	0.000	38.156	42	0.640
27	110.3	9	0.000	123.2	20	0.000	136.1	40	0.000	49.562	42	0.197
28	11.9	9	0.220	27.4	20	0.126	56.4	40	0.045	39.090	42	0.599
29	39.7	9	0.000	62.1	20	0.000	88.9	40	0.000	31.285	42	0.887
30	26.3	9	0.002	44.3	20	0.001	61.5	40	0.016	47.230	41	0.233
31	7.4	9	0.599	22.4	20	0.322	60.5	40	0.020	28.197	42	0.949
32	29.9	9	0.001	49.4	20	0.000	59	40	0.027	39.521	42	0.580
33	80.4	9	0.000	90.1	20	0.000	118.3	40	0.000	31.312	42	0.887
34	92.9	9	0.000	106	20	0.000	129.8	40	0.000	42.606	42	0.445
35	209.9	9	0.000	226	20	0.000	239.5	40	0.000	45.927	42	0.313
36	69.1	9	0.000	88.2	20	0.000	117.1	40	0.000	41.342	42	0.500
37	35.4	9	0.000	38.4	20	0.008	65.5	40	0.007	55.019	43	0.103
38	392	9	0.000	441.1	20	0.000	458.5	40	0.000	35.722	40	0.663
39	25.9	9	0.002	48.7	20	0.000	61.3	40	0.017	53.306	42	0.113
40	51.9	9	0.000	66	20	0.000	88.1	40	0.000	43.659	42	0.401
41	87.4	9	0.000	102.1	20	0.000	122.6	40	0.000	33.948	42	0.807
42	33	9	0.000	34.9	20	0.021	69.6	40	0.003	30.452	42	0.907
43	12.3	9	0.195	18.4	20	0.564	51.4	40	0.106	46.184	42	0.303
44	99.4	9	0.000	125.5	20	0.000	150.2	40	0.000	41.355	41	0.455
45	5.7	9	0.766	27.6	20	0.119	44.7	40	0.280	40.729	42	0.527
46	166.2	9	0.000	193.1	20	0.000	208.5	40	0.000	40.032	42	0.558
47	136.6	9	0.000	160.7	20	0.000	184.2	40	0.000	43.178	42	0.421
48	17.2	9	0.046	23.3	20	0.276	62.8	40	0.012	34.941	42	0.772
49	70.9	9	0.000	92.8	20	0.000	97.6	40	0.000	45.033	42	0.346
50	107.3	9	0.000	146.7	20	0.000	145.1	40	0.000	45.763	42	0.319

3PLM에 의하여 잘 설명되고 있는 것으로 나타났다. 하지만 BILOG-MG에서 b 값을 9, 20, 그리고 40 등으로 다르게 사용하였을 때 문항 적합도 검정 결과들이 약간씩 다를 수 있는 것으로 나타났다. 하지만 생성모형과 추정모형이 동일함에도 불구하고 대다수의 문항이 부적합인 것으로 나타난 결과는 동일하다고 볼 수 있기 때문에, 이러한 조건에서 어떠한 b 값을 사용하는가와 무관하게 BILOG-MG를 사용하여 문항 적합도를 살펴보는 것이 적절하지 않음을 알 수 있다.

결론적으로, 이분 문항 검사 자료 분석을 위하여 IRT를 활용하고자 하는 연구자는 문항 적합도 검정을 위하여 BILOG-MG와 같이 일반적으로 널리 쓰이는 프로그램에서 제공하는 문항 적합도 지수보다 $S\text{-}X^2$ 지수를 사용하는 것이 안전한 선택인 것으로 나타났다. $Q_1\text{-}\chi^2$ 의 경우, <표 7>에서 볼 수 있는 바와 같이 문항 수가 30이고 사례 수가 2000인 경우 1PLM과 2PLM을 사용할 때 등 매우 한정적인 조건에서만 만족스러운 통계적 검정 결과를 보이는 것으로 나타났다. 또한 표본크기가 100,000명과 같이 매우 클 때에도 무선적 표집 절차 없이 모든 피험자들의 자료를 활용하여 $S\text{-}X^2$ 지수를 계산하여도 무방하다고 볼 수 있다. 본 연구를 통하여 제시된 결과들을 바탕으로 다음과 같은 연구의 제한점 및 향우 연구 방향들을 논의할 수 있을 것으로 보인다.

첫째, <표 2>에서 제시된 바와 같이 검사 자료 생성모형에 따른 추정모형 사용 계획을 보면, 생성모형과 추정모형이 동일한 경우(GM=CM) 모의실험을 통하여 경험적 제1종 오류를 살펴보는 데에 큰 무리가 없는 것으로 보인다. 하지만 생성모형이 추정모형보다 복잡한 경우(GM>CM) 논리적으로 통계적 검정력을 살펴볼 수 있는 상황이라고 볼 수는 있으나 보다 체계적인 결론을 도출하기에는 미흡한 형태의 설계인 것으로 보인다. 왜냐하면 모의실험 연구에서의 통계적 검정력은 추정모형에 비추어보았을 때 문항 자료가 부적합인 정도와 종류에 따라서 그 결과가 다르게 나타날 수 있기 때문이다. 이는 본 연구의 주된 목적이 표본크기에 따른 제1종 오류 수준에 대한 확인에 있었기 때문에 상대적으로 소홀히 다루어진 측면이라고 볼 수 있으며, 따라서 통계적 검정력에 대한 보다 체계적인 연구 결과 및 함의를 도출하려면 표본크기와 함께 다양한 수준 및 형태의 부적합 문항 자료를 고려하는 추가적 연구가 요청된다.

둘째, 역시 <표 2>의 추정모형 사용 계획에서 비롯된 문제로서, 3PLM의 경우 본 연구를 통하여 제1종 오류 정도를 살펴볼 수 있었으나 통계적 검정력을 조사할 수 없었다. 이는 본 연구에서 사용한 세 가지 문항반응모형 중에서 3PLM이 가장 복잡한 모형이기 때문에 3PLM이 추정모형인 경우 이보다 복잡한 생성모형이 고려될 수 없었기 때문이다. 이분 문항반응 자료에 대하여 국내외의 많은 검사 프로그램들(예를 들어, 미국의 NAEP 자료 및 국내의 국가수준학업성취도 검사 자료에 대한 IRT 분석 등)에서 3PLM이 사용되고 있음을 감안할 때,

3PLM 적용 하에서의 문항 적합도 지수 수행에 대한 추가적 연구가 필요하다.

셋째, <표 1>에 대한 설명에서 확인할 수 있는 바와 같이 본 연구에서는 문항 모수의 평균과 피험자 능력모수의 평균이 일치하는 경우만을 고려하였다. 이는 검사의 난이도가 피험자 집단에 적절한 수준임을 의미한다. 현실적으로 주어진 검사가 피험자들의 능력에 비하여 매우 어렵거나 혹은 쉬울 수도 있기 때문에, 표본크기에 따른 모형 적합도 지수의 수행을 살펴보는 추후 모의실험 연구에서는 이러한 경우들에 대한 고려가 필요할 것으로 보인다.

넷째, 본 연구에서는 일차원성 가정 하에서 이분 문항자료를 다루기 위한 1PLM, 2PLM, 그리고 3PLM 세 가지 모형만을 고려하였다. 그러나, 하나의 검사에 구성형 문항이 섞여 있는 혼합 문항 검사가 대부분인 현실을 생각해 볼 때 등급반응모형(graded response model, Samejima, 1969) 등과 같이 다분 문항 자료를 다룰 수 있는 모형을 사용하는 경우에 대하여 표본크기 등에 따른 $S\text{-}X^2$ 지수의 수행을 탐구하는 추후 연구가 수행될 필요가 있다. 또한 근래 들어 활발히 연구되고 있는 다차원 문항반응모형(multidimensional item response models) 및 인지진단모형(cognitive diagnostic model) 등에 대하여 역시 다양한 크기의 피험자 수 및 문항 수 등에 따라서 $S\text{-}X^2$ 지수의 통계적 유의성 검정이 어느 정도로 양호하게 이루어질 수 있는지를 살펴볼 필요가 있다.

참고문헌

강태훈(2015). 자료크기와 집단 간 피험자 수의 차이가 차별기능문항 추출의 제1종 오류에 미치는 영향. **교육평가연구, 28**(2), 577-600.

김양분, 남궁지영, 박경호, 최인희, 강호수, 김미숙, 이규민(2016). **2016 한국교육종단연구(KELS)2013(IV): 조사개요보고서**. 기술보고 TR 2016-55-01. 한국교육개발원.

설현수(2014). Bootstrap 방법을 이용한 Rasch 이분 문항 적합도 지수의 신뢰구간 탐색. **교육평가연구, 27**(2), 279-298.

이효진, 김양수, 박 인(2013). 임상 연구에서 연구 표본수의 추출. **대한견·주관절학회지, 16**(1), 53-57.

임미경, 허 선(2007). 문항반응이론을 의사국가시험에 적용하기 위한 일차원성 및 적합도 검정. **한국의학교육, 19**(2), 163-169.

Bentler, P. M., & Bonnet, D. C. (1980). Significance Tests and Goodness of Fit in the Analysis of Covariance Structures. ***Psychological Bulletin, 88***(3), 588-606.

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

Douglas, J., & Cohen, A. S. (2001). Nonparametric item response function estimation for assessing parametric model fit. ***Applied Psychological Measurement, 25***, 234-243.

Glas, C. A. W., & Suarez-Falcon, J. C. (2003). A comparison of item fit statistics for the three-parameter logistic model. ***Applied Psychological Measurement, 27***, 87-106.

Jöoreskog, K., & Söorbom, D. (1993). ***LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language***. Chicago, IL: Scientific Software International Inc.

Kang, T. & Chen, T. T. (2008). Performance of the generalized S-X2 item fit index for polytomous IRT models. ***Journal of Educational Measurement, 45***, 391-406.

Liang, T., & Wells, C. S. (2007, April). A model fit statistic for generalized partial credit model. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. ***Applied Psychological Measurement, 9***, 49-57.

Michigan Department of Education (2011). Michigan Merit Examination Technical Manual: 2010 Testing Cycle. Retrieved from https://www.michigan.gov/documents/mde/MME_2010_Technical_Manual_final_359989_7.pdf

Nisen, J. A., & Schwertman, N. C. (2008). A simple method of computing the sample size for Chi-square test for the equality of multinomial distributions. ***Computational Statistics & Data Analysis, 52***(11), 4903-4908.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). ***IRTPRO for Windows*** [Computer software]. Lincolnwood, IL: Scientific Software International.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. ***Applied Psychological Measurement, 24***, 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. ***Applied Psychological Measurement, 27***, 289-298.

Samejima, F. (1969). ***Estimation of ability using a response pattern of graded scores***. Psychometrika Monograph, No. 17.

Sham, P. C., & Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. Retrieved from http://zzz.bwh.harvard.edu/library/statistical-power-NRG-2014-Sham-

Purcell.pdf

Sinharay, S. (2003). Bayesian item analysis for dichotomous item response theory models (ETS Research Report RR-03-34). Retrieved from http://www.ets.org/research/researcher/rr0334.htm.

Sinharay, S. (2005, April). Bayesian item fit analysis for unidimensional item response models. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. ***Journal of Educational Measurement, 37***, 58-75.

Stone, C. A., & Hansen, M. A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. ***Educational and Psychological Measurement, 60***(6), 974-991.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. ***Journal of Educational Measurement, 40***, 331-352.

Suarez-Falcon, J. C., & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. ***British Journal of Mathematical and Statistical Psychology, 56***, 127-143.

Wells, C. S. (2004). Investigation of model misfit in IRT and a new approach based on simultaneous parametric and nonparametric IRT estimation. Unpublished doctoral dissertation, University of Wisconsin-Madison.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. ***Applied Psychological Measurement, 5***, 245-262.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). ***BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*** [Computer software]. Chicago: Scientific Software International.

◎ 논문접수: 2018. 8. 15 / 수정본 접수: 2018. 9. 7 / 게재승인: 2018. 9. 15

저 자 소 개

· 강태훈 : 서울대학교 교육학과를 졸업하고 동 대학원 교육학과에서 석사학위를 취득 후, 미국 위스콘신 대학교(UW-Madison) 교육심리학과에서 양적방법론-교육측정 전공으로 박사(Ph.D.) 학위를 취득하였음. 현재 성신여자대학교 교육학과 교수로 재직 중임. 주요 관심 분야는 문항반응이론, 검사동등화, 통계적 모형의 선택 및 비교 연구, 문항 적합도 검증 그리고 인지진단모형 등임. taehoonkang@gmail.com

〈ABSTRACT〉

The Effect of Sample Size on Type I error of the S-X^2 Item Fit Index

Taehoon Kang

Sungshin University

The main purpose of this study is to examine the extent to which the $S\text{-}X^2$ item fit index that follows chi-square distributions are affected by sample sizes in terms of type I error rates through a simulation study. In addition, the performance of $Q_1\text{-}\chi^2$ index which has been traditionally used in the area of item response theory were also investigated. In the simulation study, various dichotomous item response models and a few different numbers of items were considered to check these indices could perform adequately in terms of type I error and statistical power at each simulation condition. The study results showed that the $S\text{-}X^2$ index was able to control the type I error without excessive inflation even when the sample sizes are extremely large. The $Q_1\text{-}\chi^2$ index, however, calculated by BILOG-MG program could provide the empirical type I error rates around the nominal significance level, 5%, only in very restricted conditions. In conclusion, a researcher who wants to use item response theory for the purpose of analyzing given test data needs to use the $S\text{-}X^2$ index rather than resorting to traditional item fit indices such as the $Q_1\text{-}\chi^2$ provided by commonly used commercial program like BILOG-MG. It was also shown that, if the $S\text{-}X^2$ index could be used even when the sample size was very large, then, no sampling process is required to lessen the size of data set.

Keywords : item response theory, sample size, item fit index, empirical type I error, statistical power

〔부록〕 Bilog-MG codes for 50 item conditions

1PLM	>COMMENT BILOG-MG, default prior >GLOBAL DFNAME='m11c101.dat',NPARM=1,SAVE; >SAVE PARM='CM1_101.par'; >LENGTH NITEMS=50; >INPUT NTOT=50, NID=4; >ITEMS INAMES=(SIMU1(1)SIMU50); >TEST1 TNAME=SIMU; (4A1, T1, 50A1) >CALIB NQPT=21, CYCLE=100, CRIT=0.01, CHI=(50,9); >SCORE NOPRINT;
2PLM	>COMMENT BILOG-MG, default prior >GLOBAL DFNAME='m11c101.dat',NPARM=2,SAVE; >SAVE PARM='CM2_101.par'; >LENGTH NITEMS=50; >INPUT NTOT=50, NID=4; >ITEMS INAMES=(SIMU1(1)SIMU50); >TEST1 TNAME=SIMU; (4A1, T1, 50A1) >CALIB NQPT=21, CYCLE=100, CRIT=0.01, CHI=(50,9); >SCORE NOPRINT;
3PLM	>COMMENT BILOG-MG, default prior >GLOBAL DFNAME='m11c101.dat',NPARM=3,SAVE; >SAVE PARM='CM3_101.par'; >LENGTH NITEMS=50; >INPUT NTOT=50, NID=4; >ITEMS INAMES=(SIMU1(1)SIMU50); >TEST1 TNAME=SIMU; (4A1, T1, 50A1) >CALIB NQPT=21, CYCLE=100, CRIT=0.01, CHI=(50,9); >SCORE NOPRINT;