한국심리학회지 : 일반
Korean Journal of Psychology: General
2012, Vol. 31, No. 4, 1023-1041

등급 문항 반응 모형에서 개인 합치도를 추정하기 위한 lz의 분포와 효율성

최 윤 영
한양사이버대학교

조 영 일†
성신여자대학교

개인 합치도 분석(person fit analysis)은 통계적 모형에서 예측되거나 혹은 표본 집단의 수검자들에게서 관찰된 문항 반응 형태와 다른 형태를 보이는 수검자들을 발견함으로써 개인적 수준에서 검사의 적합도를 평가하는 연구이다. 이분 문항을 위한 문항 반응 모형들에서 개인 합치도 지수 중 하나인 l_z 지수의 분포와 효과성을 조사한 연구들은 존재하지만, 다분 문항을 위한 문항 반응 모형들에서의 개인 합치도 지수에 대한 분포와 효과성를 검증한 연구는 미비하다. 본 연구에서는 등급 문항 반응 모형에서 l_z의 분포와 효율성을 조사하였다. 첫 번째 시뮬레이션 연구에서는 l_z가 다양한 조건들 하에서 표준정규분포를 보이고 있음을 입증했으며, 두 번째 시뮬레이션 연구에서는 l_z가 등급 반응 모형에서 다양한 수준의 비정상적인 반응을 보인 개인들을 탐지할 수 있음을 보여주었다. 마지막으로, l_z를 실증 자료에 적용했을 때 등급 반응 모형이 Rosenberg의 자존감 척도에 적합한 것으로 나타났다. 즉, 대부분의 개인들은 등급 반응 모형과 일치하는 문항 반응 형태들을 보였지만, 유의미한 l_z값을 가진 몇몇 참가자들은 기대했던 것과는 달리 비전형적인 문항 반응 형태를 보였다. 결론적으로, l_z가 등급 반응 모형에서 비정상적인 문항 반응 형태를 보이는 개인들을 식별해 낼 수 있는 검증력을 가졌음을 실증연구를 통해서 재확인하였다.

주요어 : 개인 합치도 지수, 표준 로그 우도 지수, 다분 문항 검사, 등급 반응 모형, 문항반응이론

† 교신저자: 조영일, 성신여자대학교 심리학과, (136-742) 서울시 성북구 보문로 34다길 2
Tel: 02-920-7593, E-mail: yicho@sungshin.ac.kr, Fax: 02-920-2040

측정은 개인의(심리적·행동적) 특성을 간단히 나타내기 위해, 그 특성에 비례하여 일정한 숫자를 부여하는 것이다(Allen & Yen, 2002). Michell(1997)은 측정을 개인의 특성을 전체 모집단에 크기에 견주어 상대적인 크기의 단위로 표시하는 것이라고 정의하였다. 검사 문항들을 통해서 개인의 특성을 점수화하거나 척도화하는 경우, 검사 점수의 질(quality)은 문항에 따른 수검자들의 반응에 측정 하고자 하는 구성 개념이 얼마나 잘 반영되었느냐에 달려있다. 또한, 검사 점수의 질은 측정 하고자 하는 구성 개념의 반영 정도 뿐 만 아니라 오류에 의해서도 크게 영향을 받는다. 여기에서 오류의 종류는 단순히 검사에 내포된 오류에서 부터 수검자의 경험이나 배경과 관련된 내용이 검사 문항에 포함되어 생기는 오류까지 다양하며, 이러한 오류들로 인해서도 수검자의 특성이 정확하게 측정될 수 없게 된다(Van der Flier, 1977). 예를 들어서, 같은 심리적 속성의 수준을 가지고 있어도 표준화에 사용된 집단과 다른 교육적·문화적 배경을 가진 수검자들은 표준화 집단과 비슷한 교육적·문화적 배경이 가진 집단의 수검자들과는 다른 관찰 점수를 가질 수 있다(Van der Flier, 1977). 이와 같이 문화적 배경과 과거의 경험에 의하여 수검자의 특성이 정확하게 측정될 수 없는 경우 뿐 만 아니라, 수검자가 검사를 받으면서 실수를 범할 경우 또는 부정행위를 하는 경우에도 낮은 수준의 개인 합치도 지수의 원인이 될 수 있다(Hulin, Drasgow, & Parsons, 1983). 예를 들어서, 추측하기는 몇 가지 문항에 대해 정확하게 혹은 부정확하게 추측함으로써 상황에 따라 각기 다른 검사점수를 개인들에게 부여한다. 그러므로 추측하기를 비롯한 다양한 시험 행동들은 검사 점수에 영향을 주기도 하고 그렇지 아니하기도 한다. 이처럼 검사점수의 질에 영향을 주는 요소들은 다양하게 존재한다.

검사의 신뢰도와 타당도는 다양한 통계 프로그램들(예, SAS, SPSS, M-Plus, 그리고 BILOG-MG)에서 제공되는 통계적 모형들(예, 고전검사이론, 요인분석, 문항반응이론 등)을 활용하여 추정할 수 있다. 하지만, 수검자의 배경이나 과거의 경험, 또는 위에 제시한 오류들로 인하여 발생하는 부정확한 측정은 전통적인 통계모형을 활용해서는 쉽게 찾아낼 수 없다. 대안으로서, 개인에게서 관찰된 문항반응형태와 준거집단에서 얻어진 것을 비교를 함으로써, 연구자는 심리적·행동적 특성 수준이 정확하게 측정되지 않은 개인들을 찾아낼 수 있는 개인 합치도 분석을 이용할 수 있다(예, Conrad, Bezruczko, Chan, Riley, Diamon, & Dennis, 2010; Emons, Sijtsma, & Meijer, 2005; Karabatsos, 2003; Sijtsma & Meijer, 2001). 개인 합치도 연구(person-fit study) 또는 개인 합치도 분석(person fit analysis)은 통계적 모형에서 예측된 개인의 문항 반응 형태나 표본 수검자 집단에게서 관찰된 문항 반응 형태를 개인들의 관찰된 문항 반응 형태와 비교함으로써 검사의 합치도를 개인적 수준에서 평가하는 통계적 방법이다. 여기에서 낮은 합치도를 보이는 반응 형태들을 “일탈된(aberrant)”, “이상한(unusual)”, “비전형적인(atypical)”, “모형에 합치하지 않은(non-model-fitting)” 반응 형태로 일컬어진다.

본 연구에서는 개인 합치도 지수들(person fit indices) 중 가장 널리 쓰이는 지수 인 l_z(표준로그 우도 지수)의 분포와 효과성을 다분 문항에서 검증하고, l_z를 실증 자료에 적용해 보았다. 다분 문항으로 구성된 검사에서 l_z,의 분

포와 효과성을 살펴보기 위해서, 등급 반응 모형(graded response model; GRM)을 이용한 두 개의 시뮬레이션 연구가 수행되었다. 첫 번째 시뮬레이션에서는 모형 적합(model-fitting) 문항 반응들이, 두 번째 시뮬레이션에서는 모형 부적합(non-model-fitting) 문항 반응들이 생성되었으며, 이 가상 자료들을 이용하여 l_z,의 분포와 효과성을 검증하였다. 실증 연구에서는 Rosenberg(1979)의 자기 존중감 척도를 사용하였다. l_z를 실증적 자료에 적용해 봄으로써, 등급 반응 모형이 자기 존중감 척도에 적합한지를 검증했다.

등급 반응 모형(Graded Response Model)

등급 반응 모형(Samejima, 1969, GRM)은 사회과학 분야에서 가장 널리 사용되는 문항 반응 이론 모형들 중 하나이다(Sung & Kang, 2006). 등급 반응 모형은 이분 문항에서 사용되는 2모수 로지스틱 모형을 서열 척도인 다분 문항을 위해서 사용할 수 있도록 적합하게 변형시킨 모형으로, 식 (1)에서 수학적으로 표현되었다(Embretson & Reise, 2000).

$$P^*_{ikj}(\theta) = \Pr(V_i \geq k \mid \theta_j, \alpha_i, \beta_{ik}) = \frac{1}{1+\exp\left[-\alpha_i\left(\theta_j - \beta_{ik}\right)\right]} \quad (1)$$

식 (1)에서 $P^*_{ik;j}(\theta)$는 무작위로 선출된 임의의 수검자 j가 i문항에서 범주 k이상의 점수를 받을 확률이다. 등급 반응 모형에서 수검자가 k라는 범주점수를 받는다는 것은 k이하의 점수인 k-1, k-2등의 범주점수는 충분히 받을 수 있다는 것을 의미한다. 식 (1)을 도표화하면 누적 문항 범주 특성곡선(cumulative item category characteristic curve)이 산출된다. 여기에서 가장 낮거나 높은 범주가 선택할 확률은 각각 1.0과 0.0으로 가정된다. 식 (1)에서 a_i 는 한 문항 안에서 모든 범주들이 동등하다고 가정되는 문항 변별 모수(item discrimination parameter)이며, β_{ik} 는 i 문항 내 k 범주의 범주 위치 모수(category location parameter)이다. 다음으로, 식 (1)을 통해 인접한(adjacent) 범주 간의 확률 차이를 근거로 하여 범주 반응 확률(category response probabilities)을 추정할 수 있으며 아래의 식과 같이 표현된다:

$$P_{ikj}(\theta) = \Pr(V_i = k \mid \theta_j, \alpha_i, \beta_{ik}) = P^*_{ikj}(\theta) - P^*_{i(k+1)j}(\theta) \quad (2)$$

인접한 누가 범주 특성 곡선들 간의 차이 값을 구하여 도표화하면 문항 범주 특성 곡선(item category characteristic curve)을 그릴 수 있다. 문항 범주 특성 곡선의 해석은 이분 문항 반응 모형의 문항 특성 곡선과 같은 방식으로, 피험자 j의 특질(또는 능력) 수준(Θ)에 따라 특정한 범주를 선택할 확률을 나타낸다.

다분 문항을 위한 개인 합치도 지수

다양한 통계적 방법들이 개인 합치도 분석에 사용될 수 있다(Meijer & Sijtsma, 2001). 그 중 기대반응을 이용하여 개인 합치도 지수를 계산하는 것에는 크게 두 가지 방식이 있다. 집단 내 다른 참가자들의 평균적 문항 반응 형태(Van der Flier, 1982)와 비교를 통해 다른 문항 반응 형태를 보이는 정도를 조사하는 방식과 문항 반응 이론 모형에 의해서 예측된

문항 반응 형태(예, Drasgow et al., 1985)와 비교하여 다른 문항 반응 형태를 보이는 정도를 추정하여 개인 합치도를 측정하는 방식이 있다. 본 연구에서는 문항 반응 모형을 기초로 하여 예측된 문항 반응에서 벗어난 형태를 보이는 정도를 지수화하여 개인 합치도 지수를 계산하는 방식을 이용하였다.

로그-우도지수(log-likelihood: l_0)는 다른 개인 합치도 지수에 비해서 많은 연구 결과들이 보고되고 있다(Nering,1995; Reise & Widaman, 1999). Drasgow (1985)은 히스토크램에 기반을 둔 l_0가 다분 문항을 가정한 문항 반응 모형들에 적용될 수 있음을 제안했다. 이후, Ro(2001)가 히스토그램 모형에 기반을 둔 l_0를 수정하여 등급 반응 모형 버전의 l_0를 소개했다. l_0의 등급 반응 모형 버전은 다음과 같이 정의될 수 있다.

$$l_0 = \sum_{i=1}^{n} \ln \left[P_{ik} \, \mathbf{V}_j \middle| \theta_j \right] \tag{3}$$

V_j는 수검자 j의 n개의 문항들에 대한 문항 반응 벡터이고, P_{ik}는 문항 i의 항목 k를 선택할 확률이다. 식 (3)에서 V_j 이외의 것은 앞에서 설명되었다. l_0는 수검자 j의 문항 반응 벡터와 θ가 주어질 경우의 관찰된 로그-우도(log-likelihood)값을 의미한다.

등급 반응 모형에서 l_0의 평균과 분산은 다음과 같이 정의된다.

$$E(l_0) = \sum_{i=1}^{n} \sum_{k=1}^{m} P_{ik}(\theta_j) \ln \left[P_{ik}(\theta_j) \right] \tag{4}$$

$$Var(l_0) = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{m} \sum_{g=1}^{m} P_{ik}(\theta_j) P_{ig}(\theta_j) \ln \left[P_{ik}(\theta_j) \right] \left[\ln \left[P_{ik}(\theta_j) \right] - \ln \left[P_{ig}(\theta_j) \right] \right] \right\} \tag{5}$$

이때 k와 g는 i 문항에 속한 범주이다. 특히, 식 (4)의 $E(l_0)$는 수검자 j의 θ가 주어질 경우에 기대되는 로그-우도값을 의미한다. 그러므로 식 (3), (4), 그리고 (5)들에 의해 등급 반응 모형을 위한 표준 로그-우도(standardized log-likelihood; l_z) 지수는 다음과 같이 정의될 수 있다.

$$l_z = \frac{l_0 - E(l_0)}{Var(l_0)^{1/2}} \tag{6}$$

덧붙여서, 다분 문항에서 사용되어지는 l_z는 이분 문항의 경우에도 쉽게 활용될 수 있다. 즉, 식 (3), (4), 그리고 (5)에서 k의 값을 2(범주의 수가 2개인 경우)로 고정함으로써, 이분 문항을 사용한 검사에서의 l_z를 추정할 있다.

기존연구들

Reise(1995)는 l_z의 영분포(null distribution)의 특징과, θ의 추정 방식에 따라 부적합한 문항 반응 형태를 찾아내는 l_z 검증력의 차이를 조사하였다. 이 연구에서 모든 θ의 추정 방식에서 l_z의 영분포는 분산은 1.0보다 작았지만, 평균은 기대값인 0.0이었다. 덧붙여서, l_z의 영분포가 비대칭적인 분포를 보인다는 것을 증명하였다. l_z의 검증력은 반가중치 추정법(biweight estimation; BIW)를 사용하는 경우에 가장 높았음을 보였다.

Meijer와 Nering(1997)은 개인의 문항 반응 형태들이 모형에서 예측한 것과 일치하지 않을 때, 발생하는 θ추정에서의 편향성을 조사하였다. 그들의 연구에서는 θ의 추정 방식으로서 최대우도법(maximum likelihood estimation; MLE), 기대된 사후 확률 추정법(expected a

posterior estimation; EAP), 그리고 BIW가 사용되었다. 그리고 다양한 θ의 추정방식 조건하에서 모형에 부적합한 문항반응벡터(non-model fitting response vectors: NFRVs)는 θ추정에 영향을 준다는 것을 발견했다. Reise와 Widaman (1999)은 모형에 부적합한 수검자들을 찾아내기 위해 고안된 문항 반응 이론과 구조 방정식 모형(SEM)에서의 개인 합치도 지수들을 비교하였다. 시뮬레이션 연구와 실증 연구를 통해서 문항 반응 이론에서 l_z값과 구조 방정식 모형에서 개인적 수준의 χ^2 값이 부적합한 문항 반응 형태들을 효과적으로 찾아낼 수 있음을 보여주었다.

대부분의 개인 합치도 지수 연구가 이분 문항을 가정한 문항 반응 모형에 중심을 두었던 것과는 달리, Ro(2001)는 다분 문항을 가정한 문항 반응 모형에서의 l_z의 분포를 검증하였다. l_z의 평균과 표준편차는 표준정규분포(N(0, 1))의 기대값과 표준편차와 같게 관찰되었으며, 이러한 특성은 θ수준 변화에 상관없이 유사한 분포를 나타냈다. 특히, (a) 문항 변별도가 중간 수준(1.0<α<1.5) 일 때, (b) 경계 난이도 분포(boundary difficulty distribution)가 θ의 분포와 동일할 때, (c) 검사가 상대적으로 길 때, l_z 분포의 평균과 표준편차가 표준정규분포의 그것들에 근접했다. 이는 기존의 이분 문항을 가정한 문항 반응 모형에서의 연구들의 결과와 일치되는 것이었다(Nering, 1995; Reise & Due, 1991). Ro의 결과들(2001)은 또한 거짓 양성 반응율(false positive rates)이 유의수준과 가깝다는 것을 밝혀 선행 연구들을 지지하였다. 그러나 Ro(2001)의 연구에서는 NFRVs를 보이는 수검자를 찾아내는 l_z의 탐지율에 검증하는 연구가 수행되지 않음으로써 l_z의 효과성을 입증하는데 한계가 있었다.

본 연구에서는 다분 문항에서 사용되는 개인 합치도 지수인 l_z의 활용성을 검증하기 위해서 다음과 같은 연구문제들을 설정하였다.

연구문제 1: 수검자의 관찰된 문항 반응 형태가 모형에서 예측된 그것과 일치될 때, l_z의 경험적 분포는 표준정규분포(standard normal distribution)를 따를 것이다.

연구문제 2: 수검자의 관찰된 문항 반응 형태가 모형에서 예측한 그것과 다를 때, l_z는 모형과 일치하지 않은 문항 반응 형태를 보이는 수검자들을 찾아낼 수 있다.

연구문제 3: 연구 참가자들의 Rosenberg 자기 존중감 척도에 대한 문항 반응 형태가 통계적으로 유의하지 않은 l_z 값을 가짐으로써, 등급 반응 모형이 척도의 문항 반응 형태를 설명하는데 적합함을 보일 것이다.

세 개의 연구문제를 검증하기 위해서, 등급 반응 모형에 기초하여 두 개의 시뮬레이션 연구들과 한 개의 실증연구가 수행되었다. 특히, 두 개의 시뮬레이션 연구들에서는, 모형과 일관된 문항 반응 형태를 보이는 수검자들로 이루어진 것과 모형과 일관되지 않은 문항 반응 형태를 보이는 수검자들이 포함된 두 개의 생성자료가 각각 사용되었다. 또한, 실증 연구에서는 l_z를 Rosenberg(1979)의 자기 존중감 척도에 적용해 봄으로써, 등급 반응 모형이 자기 존중감 척도에 적합한지를 검증했다.

연구방법

연구 1: lz 분포의 정상성

첫 번째 연구에서는 등급 반응 모형에서의

l_z의 분포를 조사하였다. GRNIRV(Baker, 1988)를 이용하여 등급 반응 모형에 기초하여 4가지의 다른 길이의 검사(10, 30, 60, 120문항)를 생성하였다. 이는, 짧은(short), 보통의(medium), 긴(long), 매우 긴(extremely long) 검사를 대표하도록 선택된 조건들이었다. 문항 난이도(β) 모수에 따른 반응을 조사하기 위해서 다음과 같은 {U(-3, 3), U(-3, 0), U(-2, 1), U(-1, 2), U(0, 3)} 5개의 문항 난이도 모수 분포들이 선택되었다. 첫 번째 조건은 잘 설계된 다중 순서화 범주(multiple ordered category)척도를 대표하는 것으로서, 이 조건에서 문항 난이도 모수는 θ의 전 범위에서 골고루 분포되었다. 두 번째 조건은 θ범위의 낮은 부분의 척도를 포함하는 것으로서, 높은 수준의 θ를 가진 피험자들의 반응 형태를 대표한다. 다섯 번째 조건은 θ범위의 높은 부분을 포함하는 것으로 이는 낮은 수준의 θ를 가진 피험자의 반응을 나타내기 위해 사용 되었다. 세 번째, 네 번째 조건들은 중간 정도 값을 나타내기 위해서 사용되었다. 다음으로, 문항 변별도 모수에 따른 l_z의 기능을 탐색하기 위해 4개의 변별 모수들(1.5, 1.00 .75, .50)이 선택되었다. 이 모수들은 다중 순서 범주 척도에서 적당한(moderate), 덜 적당한(less moderate), 낮은(low), 매우 낮은(very low) 변별도를 반영한다. 마지막으로, 1000명의 가상 피험자들이, -3.0에서 3.0 사이에서 θ 수준별로 무선적으로 산출되었다. 결론적으로, 연구 1에서는 13수준의 능력 모수(θ), 4개의 검사 길이, 5개의 범주 난이도 분포들, 4개의 문항 변별도 값들이 서로 완전히 교차되었으며, 따라서 총 1040개의 조건들이 고려되었다. 모든 조건 하에서 l_z 분포의 정규성을 검사하기 위해 평균, 표준분포, 왜도와 첨도를 계산하였으며, l_z의 1종 오류율을 검토하기 위해 거짓 양성 반응율이 계산되었다.

연구 2: lz의 부적합한 문항반응형태의 탐색

두 번째 시뮬레이션에서는 Levine과 Rubin (1979)이 제안한 거짓된 높은 점수(spuriously high; SH)과 거짓된 낮은 점수(spuriously low; SL)으로 조작하는 방법을 활용하여 부적합한 문항 반응 형태(non-fit response vectors; NFRVs)를 생성하였다. 보다 구체적으로, SH조건에서는 점수가 낮은 수검자들의 반응을 점수를 높이는 방향으로 반응들을 조작한다. 반면에, SL 조건에서는 점수가 높은 수검자들의 반응을 점수를 낮추는 방향으로 반응들을 조작한다.

첫 번째 조작은 가장 근접한 범주로 바꿈으로서 적합한 문항 반응 형태(fit response vectors; FRVs)를 부적합한 문항 반응 형태(nonfit response vectors; NFRVs)로 조작했다(예를 들면, 초기 반응이 3이었다면 SL 조건에 따라 2로, 또는 SH의 조건에 따라 4로 재입력된다). 두 번째 조작은 초기의 반응에서 두 범주를 더하거나 빼는 방식을 통해서 반응을 변화시키는 방법을 사용 했다(예를 들면, 초기 반응 3이었다면 SL 조건에서 1로, SH 조건에서 5로 재입력된다). 마지막 방법은 본래보다 세 범주를 더하거나 빼는 방식을 통해 FRVs을 생성했다(예를 들면, SL 조건에서 초기 반응 5는 2로 입력되고, SH 조건에서 초기 반응 1은 4로 입력된다).

두 번째로, 조작 비율이 l_z의 탐지율에 영향을 주는지 조사하기 위해서 3가지의 다른 조작 비율(10%, 30%, 50%)이 선택되었다. SL 조건에서는 θ수준이 0.0보다 높은 피험자의 반응들은 조작하였으며, SH 조건에서는 θ수준이 0.0보다 낮은 피험자들의 반응 벡터들을 조작

하였다. 결과적으로, 연구 2에서는 l_z의 탐지력을 검증하기 위하여 2 개의 2(조작 방향; SH vs. SL) X 3(조작 비율; 10%, 30%, 50%)과 2(조작 방향; SH vs. SL) X 3(조작 방법; 1범주 변경, 2범주 변경, 3범주 변경)의 완전무선설계가 추가되었다.

연구 3: 실증 자료에 lz의적용

연구대상

실증 연구에서는 미국의 가족전이연구(Family Transitions Project)에서 사용된 고등학교 고학년에 재학중인 542명(46.7%가 남자) 참가자의 문항 반응을 사용하였다(Conger & Conger, 2002를 참조). 표본에 대한 자세한 사항은 다른 연구에서 찾아볼 수 있다(Ge, Natsuaki, & Conger, 2006). 자료 결측에 따른 l_z의 연구의 영향을 방지하기 위해 결측치를 가진 연구 참가자의 자료는 본 연구에서 사용되지 않았다. 결과적으로, 사용된 최종 표본은 481명이였다.

연구도구

청소년의 자기 존중감 측정에 사용되는 Rosenberg(1979)의 자기 존중감 척도를 사용하였다. 이 검사는 리커트식 5점 척도인 10개의 문항들로 구성되어 있으며, 5개의 반응 선택지들은 0='절대로 그렇지 않다(strongly disagree)', 1='그렇지 않다(disagree)', 2='보통(neutral)', 3='그렇다(agree)', 4='매우 그렇다(strongly agree)'로 구성되어 있다.

결 과

첫 번째 시뮬레이션 연구를 통해서는 l_z가 고려된다양한 조건들 하에서 표준정규분포를 보이고 있음을 입증했으며, 두 번째 시뮬레이션 연구에서는 l_z가 주어진 모형(즉, 등급 반응 모형)에 기초하여 비정상적인 반응을 보이는 개인을 찾아낼 탐지력이 있음이 지지되었다. 마지막으로, 실증 자료를 l_z에 적용했을 때 등급 반응 모형은 Rosenberg의 자존감 척도에 적합한 것으로 나타났다. 대부분의 개인들은 등급 반응 모형과 일치하는 문항 반응 형태들을 보였지만, 유의미한 l_z값을 가진 몇몇 참가들은 등급 반응 모형에서 기대했던 것과는 달리 비전형적인 문항 반응 형태를 보였다. 결론적으로, l_z가 등급 반응 모형에서 비정상적인 문항 반응 형태를 보이는 개인들을 식별해 낼 수 있는 검증력을 가졌음을 실증연구를 통해서 재확인하였다.

시뮬레이션 연구 1: 모든 문항반응들이 GRM에 적합함

l_z 분포의 정규성

l_z 분포의 정규성을 확인하기 위해서 평균, 표준편차, 왜도와 첨도를 계산하고, 그 값들을 표준정규분포에 해당하는 값들과 비교했다. l_z 분포는 모든 조건에 걸쳐서 평균이 0, 표준편차가 1.00, 왜도가 -.27, 첨도가 .10인 것으로 나타나 표준정규분포의 그것들과 유사한 것으로 나타났다. 각각의 조건들에서 계산된 l_z 분포의 평균, 표준편차, 왜도, 첨도들을 표 1에 제시하였다.

l_z의 평균은 θ가 -2.0일 때 -0.1, 그리고 0.0일 때 .07까지의 범위에 있었다. l_z의 평균의 표준편차는 .03이었다. 즉, l_z는 θ의 값의 변화에 영향을 받지 않음을 알 수 있다. l_z의 표준편차의 범위는 θ가 -3.0일 때 .98, 그리고 1.5

표 1. 자료 생성 조건에 따른 l_z의 평균, 표준편차, 왜도, 첨도

Θ					문항 변별도 지수					항목 난이도 지수					검사 길이				
	M	*SD*	S^1	K^2		*M*	*SD*	S^1	K^2		*M*	*SD*	S^1	K^2		*M*	*SD*	S^1	K^2
-3.0	.00	.98	-.45	.14	1.5	.00	1.01	-.51	.31	A[3]	.00	1.00	-.37	.21	10	.00	.99	-.46	.20
-2.5	-.00	1.00	-.37	.24	1.0	.00	1.00	-.28	.12	B[4]	-.00	1.00	-.28	.08	30	.00	1.00	-.27	.10
-2.0	-.01	1.00	-.28	.12	0.75	-.00	1.00	-.16	.02	C[5]	-.00	.99	-.19	.05	60	-.00	1.00	-.20	.07
-1.5	.00	1.01	-.20	.06	0.5	.00	.99	-.12	-.03	D[6]	-.00	1.01	-.21	.07	120	-.00	1.00	-.14	.04
-1.0	.00	1.00	-.20	.02						E[7]	.00	1.00	-.29	.09					
-.5	-.00	1.00	-.16	.02															
0	.07	1.00	-.16	.30															
.5	-.00	1.00	-.17	.26															
1.0	.00	1.00	-.22	.28															
1.5	-.00	1.01	-.23	.10															
2.0	.00	.98	-.27	.10															
2.5	.00	1.01	-.36	.25															
3.0	-.00	1.00	-.45	.21															
M	-.00	1.00	-.27	.10		-.00	1.00	-.27	.10		-.00	1.00	-.27	.10		-.00	1.00	-.27	.10
SD	.03	.07	.33	.60		.03	.07	.33	.60		.03	.07	.33	.60		.03	.07	.33	.60

주. [1] S: 왜도 . [2] K: 첨도 [3] A: U(-3, 3). [4] B: U(-3, 0). [5] C: U(-2, 1). [6] D: U(-1, 2). [7] E: U(0, 3).

일 때 1.01이였으며, 이 값은 표준정규분포의 기대값과 유사했다. 그러나, θ값이 양 극단에 가까울수록 왜도의 절대값이 커졌으며(예, θ가 -3.0일 때 -0.45, 3.0일 때 -0.45), θ값이 중앙에 가까울수록 왜도의 절대값은 작아졌다(예, θ가 0.0일 때 -0.16). 첨도값 역시 비슷한 패턴을 보였는데, θ값이 양 극단에 가까울수록 첨도의 절대값은 커졌으며(예, θ가 -2.5일 때 0.24, 그리고 2.5일 때 0.25), 중앙에 가까울수록 첨도의 절대값이 작아졌다(θ가 -1.5일 때 0.02).

문항 변별도 값의 변화에 따라서는 l_z의 평균값은 0.00 그리고 표준편차의 값은 1.00에 가까운 값으로 거의 변화가 없었으며, 이는 l_z 분포가 문항 변별도 모수값에 상관없이 표준정규분포에 가깝다는 것을 의미했다. 그러나, 문항 변별도 값이 커질수록 l_z 분포는 부적으로 편포되었으며(예, α=1.5에서 -0.51, α=.5에서 -.12), 첨도 역시 문항 변별도 모수 값이 커질수록 더 커졌다(예, α=1.5에서 0.31, α=0.5에서 -0.03).

l_z 분포는 다른 조건의 문항 난이도 모수에 상관없이 표준정규분포에 가까웠다. 가장 큰 평균과 표준편차가 각각 0.00과 1.01이 나왔고, 가장 작은 평균과 표준편차가 각각 0.00과 0.99였다. U(-2, 1)가 가장 작은 왜도인 -0.19를 가졌고, 반면에 U(-3, 3)가 가장 큰 왜도인 -0.37을 가진 것으로 나타났다. 첨도에서는 U(-2, 1)가 0.05의 작은 값을 가졌고, U(-3, 3)가 가장 큰 값인 0.21을 산출해냈다.

검사 길이의 수준에 대해서는 l_z의 가장 큰 평균과 표준편차가 0.00과 1.00, 가장 작은 평균과 표준편차는 0.00과 .99인 것으로 나타나 l_z 분포가 검사 길이에 상관없이 표준정규분포와 유사함을 보여주었다. 그러나, 짧은 검사일수록 더 부적인 왜도 값을 가졌으며(예, 10 문항에서 -0.46, 120 문항에서 -0.14), 더 높은 첨도 값을 보였다(예, 10 문항에서 0.20, 120 문항에서 0.04).

거짓 양성 반응율

l_z 분포가 완전한 표준정규분포라면, 거짓 양성 반응율은 유의수준 .05에서 .05와, 유의수준 .01에서는 .01과 같아야 한다. 그러나, 대부분의 조건들에서 거짓 양성 반응율은 다소 과다 추정 되었다. 유의수준 .01에서 평균과 표준편차가 각각 .02와 .03으로 나타났고, 유의수준 .05에서는 각각 .06, .01로 나타났다. 시뮬레이션 조건에 따른 거짓 양성 반응율은 표 2에 보고되었다. θ가 1.0일 때 거짓 양성 반응율은 .026인 것을 제외하고, 극단적인 θ값은 더 높은 거짓 반응 양성율을 보였다. 예를 들어, -3.0이나 3.0과 같은 극단적인 θ수준들은 각각 .0185, .0195로 가장 높은 거짓 양성 반응율을 나타냈다. θ값 0.0은 가장 낮은 거짓 양성 반응율 .0122를 보였다.

문항 변별도 수준에 따른 거짓 양성 반응율은 문항 변별도 모수가 높을수록, 거짓 양성 반응율도 높아졌다. 예를 들면, α=0.1에서의 문항 변별도 모수값이 1.5일 때 거짓 양성 반응율은 .02였으며, 반면에 문항변별도 모수값이 0.5 일 때 거짓 양성 반응율은 .012이였다.

표 2. 자료 생성의 4가지 조건에서 l_z의 거짓 양성 반응율

Θ					문항 변별도 지수					항목 난이도 지수					검사 길이				
	α = .01		α = .05			α = .01		α = .05			α = .01		α = .05			α = .01		α = .05	
	M	*SD*	M	*SD*		*M*	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*
-3.0	.019	.012	.062	.012	1.5	.020	.010	.063	.012	A[1]	.016	.006	.060	.010	10	.018	.011	.062	.013
-2.5	.016	.007	.059	.010	1.0	.019	.061	.057	.010	B[2]	.015	.009	.058	.011	30	.019	.062	.058	.010
-2.0	.015	.007	.059	.014	0.75	.013	.006	.055	.009	C[3]	.013	.006	.056	.011	60	.014	.005	.056	.009
-1.5	.014	.006	.057	.011	0.5	.012	.005	.054	.010	D[4]	.014	.006	.056	.010	120	.013	.005	.054	.009
-1.0	.014	.005	.055	.008						E[5]	.020	.069	.057	.012					
-.5	.013	.007	.055	.011															
0	.012	.004	.055	.009															
.5	.013	.005	.054	.008															
1.0	.026	.111	.056	.009															
1.5	.014	.006	.056	.010															
2.0	.014	.006	.056	.010															
2.5	.017	.007	.060	.010															
3.0	.020	.012	.061	.011															

주. [1] A: U(-3, 3). [2] B: U(-3, 0). [3] C: U(-2, 1). [4] D: U(-1, 2). [5] E: U(0, 3).

문항 난이도 모수 수준에 따라서는 유의도 수준이 .01에서는 U(0,3)가 가장 큰 거짓 양성 반응율인 .02를 나타냈고, U(-2, 1)가 가장 작은 거짓 양성 반응율인 .01을 보였다. 마지막으로, 검사 길이의 수준에 따라서는 두 유의도 수준의 값에서 모두 검사 길이가 짧을수록 상대적으로 높은 거짓 양성 반응율을 보였다.

시뮬레이션 연구 2: 부적합한 반응에서 탐지율

시뮬레이션 연구 1에서 사용된 4가지 변인들과 부적합한(non-fitting) 반응들을 생성하기 위해 사용된 3가지 변인들에 따른 l_z의 효과성을 검증하기 위해 탐지율이 계산되었다. 표 3과 4는 각각 4개의 변인들에서의 탐지율과 조작된 3 가지의 조건들에서의 탐지율을 제시하고 있다. 극단적인 Θ값(-3.0)은 SH에서 높은 탐지율인 .75를 나타냈고, 반면에 중앙의 Θ값(0.0)은 .35의 탐지율을 보였다. SL은 가장 높은 탐지율인 .78을 Θ가 3.0일 때 생성했고, 가장 낮은 값인 .36을 Θ값이 0.0일 때 생성했다. 전반적으로, 탐지율은 Θ값이 극단으로 갈수록 증가했다.

문항 변별도 조건에 따라서는 SH과 SL 두 조건 모두에서 문항 변별도 1.5에서 .72의 탐지율을 보였고, 문항 변별도 0.5에서 탐지율은 .47이 나타났다. SL 조건에서도 1.5의 문항 변별도는 .72의 탐지율값을, 0.5의 문항 변별도에서 .47의 탐지율이 나타났다.

문항 난이도(U(-3, 3))에서 SH와 SL 조건 모두 높은 탐지율(각각 .71, .71)을 보였다. SH

표 3. 4개의 자료 생성조건에 따른 l_z의 탐지율의 평균과 표준편차

Θ					문항 변별도 지수					항목 난이도 지수					검사 길이				
	SH		SL			SH		SL			SH		SL			SH		SL	
	M	*SD*	M	*SD*		*M*	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*
-3.0	.755	.328			1.5	.716	.366	.720	.367	A[1]	.706	.339	.710	.339	10	.404	.362	.413	.363
-2.5	.733	.333			1.0	.614	.393	.617	.393	B[2]	.283	.362	.775	.316	30	.562	.387	.564	.387
-2.0	.684	.350			0.75	.533	.398	.563	.399	C[3]	.483	.392	.674	.353	60	.647	.387	.647	.388
-1.5	.602	.379			0.5	.465	.380	.473	.381	D[4]	.670	.353	.485	.394	120	.718	.375	.723	.377
-1.0	.521	.397								E[5]	.771	.319	.289	.363					
-.5	.435	.402																	
0	.350	.388	.357	.390															
.5			.433	.402															
1.0			.519	.399															
1.5			.599	.381															
2.0			.686	.349															
2.5			.736	.332															
3.0			.777	.314															

주. [1] A: U(-3, 3). [2] B: U(-3, 0). [3] C: U(-2, 1). [4] D: U(-1, 2). [5] E: U(0, 3).

조건에서 U(0, 3)는 가장 높은 탐지율인 .77을 나타냈고, 반면에 U(-3, 0)는 가장 낮은 탐지율인 .28이 나타났다. SL 조건에서는 U(-3, 0)가 .78의 가장 높은 탐지율을 보였고, U(0, -3)가 .29의 가장 낮은 탐지율을 보였다.

검사 길이수준에 따라서는 SH와 SL 조건에서 모두 긴 검사가 더 높은 탐지율을 보였다. 예를 들면, SH 조건에서 10문항을 가진 검사는 탐지율이 .40이었고, 120 문항짜리 검사는 .72의 탐지율을 보였다. SL 조건에서는 10 문항의 검사에서 .41의 탐지율을, 120 문항의 검사에서 .72의 탐지율을 가진 것으로 나타났다. 또한, 탐지율은 인위적으로 조작된 문항의 수가 증가할수록 높아졌다. 즉, 문항들 중 10%를 인위적으로 바꾸었을 때 SH에서의 탐지율은 .33, SL에서의 탐지율도 .33으로 보고되었다. 그리고 50%를 바꾸었을 때의 탐지율은 각각 SH에서 .76, SL에서 .77이 나타났다.

조작 방법에 따른 탐지율의 변화를 살펴봤을 때는 SH에서의 탐지율은 범주를 세단계 변화시켰을 때(3 범주 변경) .66, 한 단계만(1 범주 변경) 변화시켰을 때 .48이었다. 마찬가지로, SL에서는 범주를 세 단계 변화 시켰을 때의 탐지율은 .67이었고, 한 개 범주만 변화시켰을 때는 탐지율이 .48이었다.

다음으로, θ수준과 문항 변별도 수준을 함께 고려하여 l_z의 탐지율을 평가한 결과, SH와 SL 모두에서 문항 변별도 수준에 대한 탐지율은 θ가 극단으로 갈수록 증가되었다.

θ 수준과 문항 난이도 수준을 함께 고려하였을 때에 탐지율은 SH에서 문항 난이도 조건들이 U(-3, 3), U(0, 3), U(-1, 2)인 경우에는 l_z의 탐지율은 θ가 0.0에 가까워질수록 완만하게 감소하는 모습을 보였다. 그러나 문항 난이도 조건이 U(-2, 1)와 U(-3, 0)인 경우에서는 θ가 0.0에 가까워질수록 l_z의 탐지율이 0%에 가깝게 가파르게 감소하는 모습을 보였다. 반대로 SL에서는 문항 난이도 조건들이 U(-3, 3), U(-3, 0), U(-2, 1)인 경우에는, l_z의 탐지율은 θ가 0.0에 가까워질수록 완만하게 감소하는 모습을 보였다. 반면에 문항 난이도 조건들이 U(-1, 2)와 U(0, 3)일 때, θ가 0.0에 가까워질수록 탐지율이 가파르게 감소하는 모습이 나타났다.

θ 수준과 검사 길이에 따라서는 SH와 SL두 조건에서 θ가 0.0에 가까워질수록 l_z의 탐지율은 감소하는 것으로 나타났다. 마지막으로, θ 수준에 따른 문항 반응 조작 비율과 조작 방법에 따라서는 SH.와 SL 모든 두 조건에서 l_z의 탐지율은 θ가 극단으로 갈수록 증가 되는 것이 관찰되었다.

표 4. 세 개의 자료 변형 조건 하에서 l_z의 탐지율의 평균과 표준편차

조작 비율					조작 방법				
	SH		SL			SH		SL	
	M	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*
10%	.330	.304	.334	.311	1 범주 변형	.479	.371	.483	.370
30%	.654	.378	.655	.378	2 범주 변형	.604	.394	.609	.396
50%	.763	.365	.772	.358	3 범주 변형	.665	.399	.669	.397

실증연구: Rosenberg의 자존감 척도에 lz의 적용

Rosenberg의 자존감 척도의 모형 검증

Rosenberg(1965, 1979)의 자존감 척도에서 평균과 표준편차를 계산하였다. 소척도 내 문항들의 평균은 모두 2.5 보다 컸으며, 이는 청소년 대부분이 높은 수준의 자존감을 지니고 있는 것을 의미했다. 달리 말하면, 많은 참가자들이 대부분의 문항에서 "그렇다" 또는 "매우 그렇다"로 응답했다는 것을 의미했다. 분석하기 전에, 문항반응이론의 가정 중의 하나인 일차원성(unidimensionality)를 검증하기 위해서 SAS 9.1을 이용해 탐색적 요인분석(EFA)을 시행했다. 일반적으로 이 척도는 오직 청소년의 자존감 요인 하나만을 측정하기 위해서 사용되기 때문에 요인 구조를 회전시키기 않았다. 첫 번째와 두 번째 요인에 의해 설명되는 분산은 각각 5.12와 .79였다. 따라서 Rosenberg의 자존감 척도의 소척도 문항들의 일차원성 가정은 EFA를 통해 지지되었다.

Rosenberg의 자존감 척도에서 개인 합치도

l_z를 산출하기 위해 문항 모수와 개인 모수를 MULTILOG를 통해 추정했다. 10문항에서 추정된 문항 모수치는 부록 A에 보고되어 있다. 전반적으로, 상대적으로 낮은 범주 난이도가 추정되었으며, 상대적으로 높은 값의 문항 변별도 값이 추정되었다. 범주 난이도에서의 적은 값은 "그렇다" 범주와 "매우 그렇다" 범주가 선택되기 위해서 중간 수준의 자존감을 필요로 한다는 것을 의미하며, 문항 변별도의 큰 값이 의미하는 바는 문항들이 자존감 수준이 높은 청소년과 상대적으로 자존감 수준이 낮은 청소년들을 잘 구별한다는 것을 의미한다.

또한, 자존감에 대한 잠재 변인의 요인 점수를 추정하여, 이 점수와, l_0, l_z의 상관계수를 계산하였다. 자존감의 요인점수와 l_0, l_z와의 상관값은 각각 .589(p<0.01), .095(p=0.3)로 나타났다. 따라서 l_z는 상대적으로 l_0보다 자존감의 수준에 영향을 덜 받았다.

l_z의 분포에서, l_z의 평균과 표준편차는 각각 0.075, 0.686으로 나타났으며, 왜도는 -2.72, 첨도는 11.06으로 추정되었다. l_z의 분포를 도표로 보기 위해서 l_z의 히스토그램은 그림 1에 제시하였다.

Rosenberg 척도에서 l_z의 탐지율은 약 3% (n=15)의 청소년에게서 -1.645보다 적은 l_z가 관찰되었다. 3%는 단측 검정의 1종 오류(5%)보다 낮은 비율이기 때문에, Rosenberg의 자존감 척도가 GRM에 적합함을 의미한다. l_z가 개인 합치도 지수로써의 활용을 설명하기 위해서 표 5에 등급 반응 모형과 일관된 문항 반응 형태를 보인 두 명과 일관되지 않은 문항 반응 형태를 보인 두 명의 수검자들의 그것들을 예로서 제시했다. 표 5의 상단 부분에는 l_z의 값이 가장 작았던 두 명, 즉 l_z 지수가 개

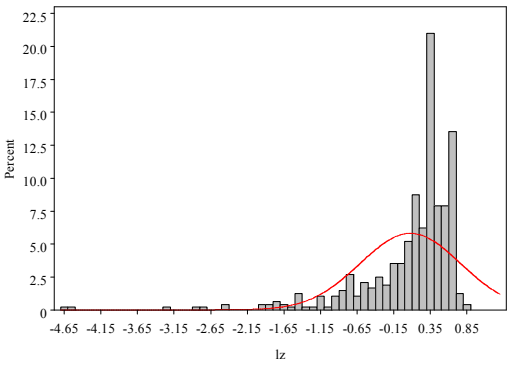

그림 1. Rosenberg의 자기존중감 척도에서의 l_z의 분포

표 5. l_z값에 따른 수검자들의 예측된 문항반응형태와 관찰된 문항반응형태

문항	관찰값	ID = 440 (l_z = -4.51 and Θ = -2.47) 범주별 예측 확률					관찰값	ID = 301 (l_z = -4.64 and Θ = -1.41) 범주별 예측 확률				
		0	1	2	3	4		0	1	2	3	4
1	4	0.19	0.32	**0.39**	0.10	0.00	4	0.01	0.04	0.26	**0.67**	0.02
2	4	0.10	0.29	**0.48**	0.12	0.00	4	0.01	0.03	0.26	**0.69**	0.02
3	0	**0.45**	0.19	0.30	0.05	0.00	0	0.02	0.03	0.29	**0.62**	0.04
4	4	0.18	**0.45**	0.30	0.07	0.00	4	0.01	0.08	0.35	**0.54**	0.01
5	0	**0.41**	0.26	0.23	0.10	0.00	4	0.06	0.09	0.29	**0.51**	0.05
6	0	**0.44**	0.46	0.09	0.01	0.00	0	0.02	0.16	**0.63**	0.18	0.00
7	0	0.34	**0.50**	0.15	0.02	0.00	0	0.02	0.18	**0.55**	0.25	0.00
8	0	0.41	**0.48**	0.08	0.03	0.00	4	0.07	**0.40**	0.32	0.20	0.02
9	0	**0.56**	0.36	0.06	0.02	0.00	0	0.15	**0.46**	0.25	0.13	0.01
10	0	0.43	**0.50**	0.06	0.02	0.00	0	0.04	**0.37**	0.33	0.24	0.01
		ID = 363 (l_z = .80 and Θ = -1.95)						ID = 440 (l_z = .82 and Θ = -1.97)				
1	2	0.05	0.14	**0.47**	0.33	0.00	2	0.05	0.15	**0.48**	0.32	0.00
2	2	0.03	0.11	**0.50**	0.35	0.00	2	0.03	0.11	**0.51**	0.34	0.00
3	2	0.13	0.12	**0.51**	0.24	0.01	2	0.14	0.12	**0.50**	0.23	0.01
4	2	0.05	0.25	**0.47**	0.23	0.00	3	0.06	0.26	**0.46**	0.22	0.00
5	2	0.17	0.20	**0.35**	0.26	0.01	2	0.18	0.21	**0.35**	0.25	0.01
6	1	0.11	**0.49**	0.37	0.03	0.00	1	0.12	**0.49**	0.36	0.03	0.00
7	2	0.10	**0.43**	0.40	0.07	0.00	1	0.11	**0.44**	0.39	0.06	0.00
8	0	0.19	**0.54**	0.19	0.08	0.01	1	0.20	**0.54**	0.19	0.07	0.01
9	1	0.33	**0.48**	0.13	0.05	0.00	0	**0.33**	0.48	0.13	0.05	0.00
10	1	0.15	**0.60**	0.18	0.07	0.00	1	0.16	**0.60**	0.17	0.07	0.00

주: 0='절대로 그렇지 않다', 1='그렇지 않다', 2='보통(neutral)', 3='그렇다', 4='매우 그렇다'.

인 합치도 지수가 낮다고 식별한 이들의 문항 반응형태를 나타내고 있다. 식별번호 440인 청소년에게서 관찰된 반응 형태를 살펴보면 10문항 중 4문항만이 기대확률이 높은 범주와 관찰된 범주와 일치할 뿐이었다. 식별번호가 301인 청소년은 단 한 문항도 관찰된 반응과 기대된 반응이 일치되는 것이 없었다. 반대로, 표 5의 하단 부분은 FRVs를 보인다고 가정되는 두 명의 관찰된 반응과 범주의 기대확률이 제시되어 있다.

우선, l_z가 0.8인 식별번호가 363인 청소년이 보인 실제 반응들 중 8개가 각 문항에서 기대확률이 높은 범주와 일치되었음을 알 수 있다. l_z가 .82인 식별번호가 440인 청소년은

자존감 수준에서 기대확률값이 가장 큰 범주와 관찰값을 비교하였을 때 9개의 문항이 일치함을 알 수 있다.

논 의

본 연구에서는 다분 문항에서 사용되는 개인 합치도 지수인 l_z의 활용성 검증하기 위해서 세 개의 연구 문제들을 설정하고 시뮬레이션과 실증 연구를 통하여 이것들을 검증하였다.

첫 번째 시뮬레이션 연구를 통하여 수검자의 관찰된 문항 반응 형태가 모형에서 예측된 그것과 일치될 때, l_z의 경험적 분포는 θ, 문항 난이도, 문항 변별도의 수준과 검사 길이에 상관없이 표준정규분포(standard normal distribution)를 따름을 보였다. 이는 l_z가 등급반응 모형에 적합하지 않은 반응들을 탐지해 내는 합치도 지수로써 사용될 수 있음을 보여주었다. 이러한 결과는 이분 문항일 때(Nering, 1995)와 다분 문항 조건에 근거한 선행 연구들(Ro, 2001)과 일치하였다.

두 번째 시뮬레이션 연구를 통하여 수검자의 관찰된 문항 반응 형태가 모형에서 예측한 그것과 다들 때, l_z는 모형에 적합하지 않은 문항 반응 형태를 보이는 수검자들을 효과적으로 찾아낼 수 있음을 검증하였다. 특히, 탐지율은 검사가 길어지고, 보다 많은 수의 항목들이 조작되고, 더 많은 문항들이 조작될수록 높아졌다. 이것은 이분 문항을 다룬 선행 연구들(Nering, 1995; Reise, 1995)과 유사한 결과였다. 특히, 문항 난이도 모수가 θ와 일치하지 않을 때는 그렇지 않을 때보다 탐지율이 높았다. 또한, 연구 2에서는 조작 방식, 조작 비율, SH/SL에 따른 l_z의 탐지율의 변화를 추가로 관찰하였다. 문항 난이도 모수의 분포와 조작 비율이 l_z 의 탐지율에 다른 조작 변수들에 비하여 상대적으로 많은 영향을 미쳤으며, 이는 θ가 문항 난이도 모수 분포로부터 멀리 있을 때 비전형적인 문항 반응 형태를 보이는 수검자를 찾는데 보다 효과적임을 의미한다. 결론적으로, 연구자가 NFRVs를 탐지하고자 할 때, l_z는 문항 난이도 모수 분포가 θ와 많은 차이가 있는 문항에서 효과적임을 유추할 수 있다.

마지막으로 경험적 사례연구를 통해서 l_z가 어떻게 비전형적인 문항 반응 형태를 보이는 개인을 찾아내는데 사용될 수 있는지를 보여주었다. 본 연구에서는 l_z를 Rosenberg(1979)의 자기 존중감 척도에 적용해 봄으로써, 등급반응모형이 자기 존중감 척도에 적합한지를 검증하였다. 연구참여자들 중의 3%만이 유의미한 l_z를 보였기 때문에 등급 반응 모형은 Rosenberg의 자존감 척도의 문항 반응을 잘 설명하고 있음을 확인하였으며, 이는 나아가 l_Z가 모형의 동일성 검증 연구에 활용될 수 있는 가능성을 시사하였다(Van der Filer, 1977).

하지만, 본 연구는 다음과 같은 몇 가지 제한점을 가지고 있다. 비록 실증 연구를 통해서 시뮬레이션 자료를 이용한 분석 결과를 뒷받침했지만, 두 시뮬레이션 연구들에서 l_Z의 계산에 사용된 θ는 추정된 값이 아니라 진점수 θ에 근거하고 있다는 것이다. 따라서 후속 연구에서는 진점수 대신에 추정된 θ값을 사용한 연구를 할 필요성이 있다.

다음으로, 본 연구에서는 거짓 양성 반응율을 계산할 때 양측 검증 대신 단측 검증을 사용하였다. Nering (1995)은 단측 검증보다 양측 검증이 l_z의 분포를 검증하는데 보다 효과적일

수 있음을 주장했다. 이에 근거하여 거짓 양성 반응율의 추정에서 양측 검증을 사용한 조사를 해볼 필요성이 있다.

마지막으로, 시뮬레이션에서의 자료는 균등분포(uniform distribution)를 가정한 항목 난이도 모수 분포에 근거하여 생성되었다. 그러나, 이 가정은 다분 문항을 사용하는 심리 검사들의 항목 난이도 모수 분포가 균등분포에 근거하지 않고 있다는 현실과 일관되지 못하는 제한점을 가지고 있다. 그러므로, 항목 난이도 모수가 균등 분포 외에 정규 분포를 따르는 가상 자료를 이용한 후속 연구를 해볼 필요성이 있다.

이러한 한계점들에도 불구하고, 본 연구는 다분 문항을 위한 등급 반응 모형에서 개인 합치도 지수로써의 l_z의 유용성을 시뮬레이션 연구과 실증 연구를 통해서 보여주었다는 점에서 중요한 시사점을 준다. 보다 구체적으로, 개인 합치도 지수인 l_z를 활용하여 통계 모형의 적합도를 평가하는 경우에는 일정 수(1% 또는 5%) 이상의 수검자들이 유의미한 l_z를 가지는 경우에는 그들을 하나의 하위 집단으로 보고 이들의 심리적 특성을 따로 분석하는 것이 보다 타당해 보인다. 그리고 잠재계층모형(latent class models)을 사용하기에 앞서서 개인 적합도 지수를 활용하여 사전연구를 수행해 볼 수 있을 것이다.

참고문헌

Allen, M. J., & Yen, W. M. (2002). *Introduction to Measurement Theory*. Prospect Heights, IL: Waveland.

Baker, F. B. (1988) *GENIRV: A computer program for generating item response*. Unpublished manuscript, University of Wisconsin, Madison.

Conger, R. D. & Conger, K. J. (2002). Resilience in Midwestern families: Selected findings from the first decade of a prospective, longitudinal study. *Journal of Marriage and Family*, *64(2),* 361-373.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response model and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38(1)*, 67-86.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoterical sampling distributions of the *U3* person-fit statistic. *Applied Psychological Measurement. 26(1)*, 88-108.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, Local, and Graphical Person-Fit Analysis Using Person-Response Functions. *Psychological Methods*, *10(1),* 101-119.

Ge, X., Natsuaki, M. N., & Conger, R. D. (2006). Trajectories of depressive symptoms and stressful life events among male and female adolescents in divorced and nondivorced families. *Development and Psychopathology*, *18*(1), 253-273.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory, *Applied Psychological*

Measurement, 20, 101-125.

Karabatsos, G. (2003). Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics. *Applied Measurement in Education*, *16(4)*, 277-298.

Levine, M, V., & Rubin, D. F. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4(4)*, 269-290.

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement*, *21(4)*, 321-336.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, *8(3)*, 261-272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology Review: Evaluating Person Fit. *Applied Psychological Measurement*, *25(2)*, 107-135.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88(3)*, *355-383*.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, *21(4)*, 121-129.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, *19(3)*, 213-229.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics in the detection of aberrant response patterns. *Applied Psychological Measurement*, *15*, 217-226.

Reise, S. P., & Widaman, K. W. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods*, *4*, 3-21.

Ro, S. (2001). *Characteristics of a likelihood-based person-fit index under the graded response model*, Doctorial dissertation, University of Minnesota, Minneapolis.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement Number 17*, 34(4), Part 2.

Schmitt, N., Cortina, J. M., & Whitney, J. M. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement*, *17*, 143-150.

Sung, H. J., & Kang, T. (2006). *Choosing a polytomous IRT model using Bayesian model selection methods*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Trabin, T. E., & Weiss, D. J. (1979). *The person response curve: Fit of individuals to item characteristic curve models*. Research Report 79-7, Computerized Adaptive Testing Laboratory, Department of Psychology, University of Minnessota, Minneapolis.

Van der Flier, H. (1977). Environmental factors and deviant response patterns. In Y.H. Pootinga (Ed.), *Basic problems in cross-cultural psychology*. Swets & Seitlinger, B. V., Amsterdam.

Van der Flier, H. (1982). Deviant response patterns and comparability of test scores.

Journal of Cross-Cultural Psychological, *13(3)*, 267-298.

1 차원고접수 : 2012. 6. 30.
수정원고접수 : 2012. 10. 3.
최종게재결정 : 2012. 10. 29.

한국심리학회지 : 일반
Korean Journal of Psychology: General
2012, Vol. 31, No. 4, 1023-1041

An Evaluation of lz as a Measure of Person Fit Under Graded Response Models

Younyoung Choi
Department of Counseling Psychology
Hanyang Cyber University

Young Il Cho
Department of Psychology
Sungshin Women's University

A person-fit analysis (PFA) has been developed to identify individuals whose latent traits are not measured accurately by a test. This study investigated the distribution and effectiveness of a person-fit index: the standardized log-likelihood index (l_z) for Graded Response Model(GRM). Findings in simulation studies employing various manipulated variables demonstrated that the empirical distribution of the l_z were close to the standard normal distribution and the l_z was effective in detecting individuals showing aberrant response patterns given the IRT model (i.e., graded response models). The application of the l_z to the empirical data of the Rosenberg Self-Esteem scale (1965, 1979) illustrated that GRM was suitable to the scale, producing relatively few individuals of a significantly large value.

Key words : Person-fit index, Standardized log-likelihood index, Polytomous items, Aberrant testing behavior, Rosenberg self-esteem scale

부록 A

Rosenberg 자기존중감 척도의 문항 변별도 및 범주 난이도 추정치

Items	a	b_1	b_2	b_3	b_4
나는 내가 다른 사람들처럼 가치 있는 사람이라고 생각한다.	2.77 (.24)	-3.00 (.40)	-2.46 (.23)	-1.70 (.13)	0.07 (.07)
나는 좋은 성품을 가졌다고 생각한다.	2.72 (.25)	-3.27 (.53)	-2.63 (.26)	-1.74 (.14)	0.00 (.07)
*나는 대체적으로 실패한 사람이라는 느낌이 든다.	3.30 (.36)	-2.53 (.22)	-2.29 (.19)	-1.62 (.11)	-0.45 (.06)
나는 대부분의 다른 사람들처럼 일을 잘 할 수 있다.	2.65 (.23)	-3.04 (.43)	-2.27 (.21)	-1.50 (.11)	0.26 (.08)
*나는 자랑할 것이 별로 없다.	2.27 (.23)	-2.64 (.35)	-2.18 (.22)	-1.52 (.14)	-0.10 (.08)
나는 내 자신에 대하여 긍정적인 태도를 가지고 있다.	3.53 (.33)	-2.54 (.23)	-1.84 (.12)	-1.00 (.07)	0.51 (.06)
나는 내 자신에 대하여 대체로 만족한다.	2.86 (.25)	-2.71 (.30)	-1.90 (.14)	-1.04 (.08)	0.74 (.08)
*나는 가끔 내 자신이 쓸모없는 사람이라는 느낌이 든다.	2.08 (.19)	-2.65 (.29)	-1.48 (.13)	-0.79 (.09)	0.52 (.09)
*나는 내 자신을 좀 더 존경할 수 있었으면 좋겠다.	1.83 (.17)	-2.35 (.24)	-1.16 (.12)	-0.41 (.09)	1.03 (.12)
*나는 때때로 내가 좋지 않은 사람이라고 생각한다.	2.72 (.24)	-2.58 (.24)	-1.55 (.11)	-1.02 (.09)	0.21 (.07)

주: 괄호안의 숫자는 표준오차를 의미한다.
*는 역문항을 의미한다.