선진국방연구
Journal of Advances in Military Studies
2019, Vol. 2, No. 1, 37-55.
https://doi.org/10.37944/jams.v2i1.41

군 리더 진성 리더십을 어떤 문항을 사용하여 측정할 것인가?

배성호*
(광운대학교)

≪국문초록≫

진성 리더십을 정밀히 측정하는 연구는 활발하지 못하며, 특히 진성 리더십을 측정하여 어떠한 의사결정을 내리고자 할 때는 개별 문항의 가동 범위와 신뢰도가 상세히 알려져 있지 못하다. 이 연구는 군 현장 리더가 자신의 상급 리더를 평가한 자료를 토대로 다층 다수준 문항 반응 모형을 분석해 군 리더 진성 리더십을 적절히 측정할 수 있는 다섯 문항을 제시하였다. 그 문항은 8번, 12번, 13번, 14번 16번으로 나타났다. 또한 이들 문항은 진성 리더십이 낮은 리더를 변별하는 일에 유용하게 쓰일 수 있음을 알게 되었다. 향후 과제로는 쌍요인(Bifactor) 구조를 검토하는 일이 있으며, 진성 리더십이 군 인사 결정에 어떤 방향으로 필요한지 충분한 토론을 해야 하는 점을 제시하였다.

주제어 : 문항반응이론, 적절한 측정, 가동 범위, 문항 선택

* me@seonghobae.me

Ⅰ. 소개

어떤 조직이든 리더를 선발할 때 인성 검사가 중요하게 기능할 필요가 있다(Carnes, Houghton, & Ellison, 2015). 지능을 측정하는 '인지능력검사'는 군 리더 선발 검사에서 일정 수준 이상이면 통과하는 방식으로 변화하였다. 그래서 인성 검사(성격 검사)와 성격 검사 결과를 토대로 진행한 면접시험의 정확성이 향후 군 리더 선발 상황에서 중요하다(Kwak & Choi, 2016; Lee & Park, 2017). 군 장면(military setting)에서 사용하는 심리 검사는 검사 결과가 승진 결정이나 선발 결정, 교육 결정에 중대한 영향을 줄 여지가 있기 때문에 문장을 이해하지 못해 발생하는 검사 측정 오류 또는 사회적으로 바람직한 내용(socially desirable contents)을 담아 발생하는 검사 측정 오류가 발생할 수 있는 심리 검사 문항을 제거하여 문항을 신중히 선택해 사용해야 한다. 그런데 이미 출판된 심리 검사(psychological testing)이더라도 주성분 분해(Principal Component Factoring)를 이용한 결과를 사용하거나 측정 변수의 측정 수준을 서열형이 아닌 연속형으로 취급하는 경우(Rhemtulla, Brosseau-Liard, & Savalei, 2012a)가 언제든 있을 수 있다(Henson & Roberts, 2006). 이렇게 주성분 분해와 잘못된 측정 수준 명세는 측정 오차(measurement error)를 만들어 내고, 측정 오차를 무시하는 결과는 부정확한 인성 검사를 만들게 되어 적절한 리더를 선발하지 못하는 결과를 얻는다. 게다가 주성분 분해와 잘못된 측정 수준 명세가 이루어지면 어떤 문항이 사회적으로 바람직한 내용을 담고 있는지 알 수 없기 때문에(Holden & Book, 2009), 리더 선발용 인성 검사 문항으로서 중요한지 여부를 알기 어렵다.

덧붙여서 기존에 출판된 심리 검사에는 동일 방법 효과(common method effect)가 포함돼 있다. 처음에는 자기 보고(self-report)에 따른 문제로만 치부됐었다(Podsakoff & Organ, 1986). 그런데 실제로는 주성분 분석과 상관 분석 관행에 따른 "유사한 문장(또는 단어) 효과" 때문에 방법 효과가 생길 수 있음(Podsakoff, MacKenzie, Lee, & Podsakoff, 2003)을 알게 되었다.

이런 효과를 제어하기 위한 전략이 몇 가지 있다. 탐색 구조방정식 모형(ESEM: Exploratory Structural Equation Modeling)을 이용하는 방법(Ferrando, 2009; Marsh et al., 2010; Marsh, Nagengast, & Morin, 2013; Morin, Katrin Arens, & Marsh, 2015), 그리고 문항반응이론(IRT: Item response theory)를 이용하는 방법(Bock, 1997; Brown, Inceoglu, & Lin, 2017; Cai, 2008; Cao, Drasgow, & Cho, 2015; Carter et al., 2014; Kang & Chen, 2008; LaHuis, Clark, & O 'Brien, 2011; Reise & Flannery, 1996; Sinharay, 2015; Tendeiro, 2017; Thissen, Cai, & Bock, 2010)이다. 이 중 문항반응이론을 이용하여 군 리더 진성 리더십을 평가할 수 있는 오차가 적은 문항을 선택할 것이다.

1.1 공통 방법 편향 문제와 해결 전략

1.1.1 구조방정식모형 접근과 문항반응이론 접근

구조방정식모형 접근과 문항반응이론 접근 간 가장 큰 차이는 문항 오차 간 상관을 허용(Kline, 2011, pp. 251-252)하는지 아닌지(Allen & Yen, 1979)에 달렸다. 구조방정식모형 접근은 CTCU (Correlated Trait Correlated Uniqueness) 접근을 이용하여 검사 문항 간 내용이 유사하더라도 문항이 적절한 것으로 판정(Marsh et al., 2013)하는 경향이 있다. 반면 문항반응이론은 측정 오차 간 상관을 허용하지 않음으로써 문항의 가능도 계산(Likelihood calculation) 때 기대 가능도 $E(LL)$과 관찰된 LL 간 격차가 큰 것으로 판정해 공통 방법 편향이 발생해 문항 오차 상관이 발생하는 문항을 제거한다(Chalmers & Ng, 2017; Drasgow, Levine, & Williams, 1985; Kang & Chen, 2008, 2011; Liu & Chalmers, 2018; Tay & Drasgow, 2012; Tendeiro, 2017). 이로서 결과적으로는 모든 내용을 포괄하면서 상호 독립적인 문항 구성을 얻게 된다.

1.1.2 문항반응이론이 우수한 이유: 왜곡 응답(faking response) 검출

인사 선발 결정을 위한 심리 평가는 고부담 검사(high-stakes testing)라 칭한다(Attali, Lewis, & Steier, 2013; Levashina, Weekley, Roulin, & Hauck, 2014; Makransky & Glas, 2013). 고부담 검사는 항시 왜곡 응답(Griffith, Chmielowski, & Yoshita, 2007)을 수반한다. 그런데 이 때 유사한 응답 패턴이 반복되는 경우 문항 오차 간 상관이 발생(Podsakoff et al., 2003)하여 공통 방법 편향이 나타날 수 있으므로 문항 오차 상관이 발생하는 문항을 제거할 수 있게 된다(Chalmers & Ng, 2017; Drasgow et al., 1985; Kang & Chen, 2008, 2011; Liu & Chalmers, 2018; Tay & Drasgow, 2012; Tendeiro, 2017).

1.2 진성 리더십 검사

1.2.1 초기 개발

심리 검사를 만들 땐 무엇을 측정할지 명확한 정의(개념 준거: conceptual criterion)가 있어야 실제 행동(실제 준거: behavioral criterion)을 토대로 측정이 가능하다(Campbell, Dunnette, Arvey, & Hellervik, 1973). 그런데 진성 리더십(authentic leadership)은 성공적인 진성 리더의 핵심 사건(critical incidents)이 무엇인지 명확하지 못하다. 진성 리더십이 검사 문항으로 개발되기 이전 시기에는 “깨어 있고”(self-aware), “공명정대하며”(unbiased processing), “진심으로 행동하고”(Authentic Behavior and actions), “진실한 관계를 맺는”(relational authenticity) 리더의 특성으로 주목 받았다

(Ilies, Morgeson, & Nahrgang, 2005). 그 이후 세대에 진성 리더십 측정 도구가 처음 개발될 때(Walumbwa, Avolio, Gardner, Wernsing, & Peterson, 2008) 진성 리더십 측정 도구는 핵심 구심점 없이 모든 선행 연구를 집약해 버리는 오류를 범하였다. 본래 심리 검사를 개발할 때는 모든 좋은 점을 포괄하기 보다는 "타인과 비교할 수 있는 핵심 사건"을 중심으로 "행동"(behavior)을 측정해야 한다(Campbell et al., 1973). 그럼에도 Walumbwa와 동료(2008)는 Campbell과 동료(1973, p. 16)가 수행한 "무엇이 (진성 리더십에서) 가장 의미 있고 중요한 요소인가?"를 논의하는 절차를 생략해버렸다. 그 결과 "무엇을 잴 것인가"에 해당하는 '개념 준거'의 정의가 지나치게 길고 모호해졌다. "a pattern of leader behavior that draws upon and promotes both positive psychological capacities and a positive ethical climate, to foster greater self-awareness, an internalized moral perspective, balanced processing of information, and relational transparency on the part of leaders working with followers, fostering positive self-development"(Walumbwa et al., 2008, p. 94)라는 다소 길고 불명확한 측정적 정의를 갖게 됐다. 결과로는 단 16문항으로 너무 넓은 영역을 측정하려 하게 됐고, 이는 문항 내용이 무엇을 묻고 있는지 명확하지 않게 보이는 결과를 낳았다.

1.2.2 요인 구조

진성 리더십 검사(Walumbwa et al., 2008) 개발 절차가 불명확한 면이 있으나, 총 16문항으로 구성돼 있고, 이들 개념은 고차 요인 구조(second-ordered factor structure)로 알려져 있어 세부 요인과는 무관하게 '진성 리더십'이란 큰 개념으로 사용할 수 있다. 심지어는 Walumbwa와 동료(2008)가 주장한 4요인 구조를 무시하고 16문항 전체를 한 요인으로 처리하더라도 CFI(Comparative Fit Index)와 RMSEA(Root Mean Square Error of Approximation) 보고가 소규모 표본에서는 받아들일 수도 있는 수준으로 나타났기 때문에(Kenny, Kaniskan, & McCoach, 2015; Lai & Green, 2016) 한 요인으로 처리하는 것도 무관하다. 왜냐하면 Walumbwa와 동료 (2008)가 약 200명을 표본으로 하여 검사를 만들었는데, 이 때 CFI가 .91로 좋으나 RMSEA는 약 .10에 달하여 1요인 구조를 받아들이지 못하는 걸로 오해하였다(Hu & Bentler, 1999). 그런데 Hu와 Bentler의 시뮬레이션과는 달리 Kenny, Kaniskan과 McCoach (2015) 시뮬레이션 결과에서는 200명 표본일 때 RMSEA가 권장 값인 .05보다 클 확률은 최대 .387에 달한다. 이 때문에 Hu와 Bentler의 시뮬레이션 결과가 절대 기준이 될 수는 없으며, CFI가 .9가 넘지만 RMSEA가 .10을 넘겨 상황이 좋지 못할 때(Lai & Green, 2016)는 RMSEA에 의한 판단을 보류할 필요가 있다.

게다가 세부 내용 영역으로서 요인 구조는 'Self-Awareness', 'Relational Transparency', 'Internalized Moral Perspective', 'Balanced Processing'으로 구성돼 있다(Walumbwa et al., 2008, p. 99). 그럼에도 이들은 고차 요인 구조이기 때문에 문항반응이론에서는 세부 내용 영역이 국면(facet) 요인일 뿐이어서(Makransky, Mortensen, & Glas, 2013) '진성 리더십'이란 큰 맥락 아래에서 어떤 검사 문항을 사용하든 큰 문제는 없을 걸로 볼 수 있다.

1.2.3 비평

진성 리더십 검사 연구는 분명 검사 개발 연구(Walumbwa et al., 2008)임에도 검사 특성 추정치 수치가 잘 드러나 있지 않다. 구조방정식 모형을 사용하여 검사를 개발하였음에도 공통방법요인이 있는지, 오차 간 상관이 발생하는지, 어떤 문항이 우수한 문항인지 언급이 없으며, 부록에 제시한 예제 문항(Walumbwa et al., 2008, p. 121)조차 선택한 이유를 언급하고 있지 못하다. 상황이 이렇기에 군 리더의 진성 리더십을 측정하고자 한다면 반드시 우수한 문항이 무엇인지, 측정 오차가 적은 문항이 무엇인지 규명하고 사용하여야 측정 품질을 담보할 수 있다.

Ⅱ. 연구 방법

2.1 자료 및 평가 방법

자료는 진성 리더십 선행 연구(Yang, Kim, & Kim, 2017)에서 사용한 자료를 입수해 이용하였다. 동료 평가 자료(peer rating data)를 사용하였다. 302명이 평가 받았고, 203명이 평가하였다. 소대장이 중대장을, 중대장이 대대장을, 소대장이 대대장을 평가하는 방식으로 평가되었다. 따라서 소대장 개인 ID와 중대장 개인 ID, 대대장 개인 ID는 데이터에 식별 가능토록 처리돼 있다.

2.2 연산

2.2.1 추정 소프트웨어

문항반응이론 연산은 자동 탐색적 요인 분석 소프트웨어(Bae, 2017)를 이용하였다. 추정 절차에는 통제 변수(control variables)로서 지위 (대대장, 중대장, 소대장), 평가자 고유 식별 코드, 피평가자 고유 식별 코드를 사용하였다. 그리고 이들 변수 조합을 생성하여 상호작용까지 고려하였다. 다만 이 중 해가 수렴하지 못하였거나 표준오차 계산이 효율적이지 못한 모형은 채택 대상에서 제외됐다. AIC(Preacher, Zhang, Kim, & Mels, 2013)를 기준으로 최종 해가 선택됐다.

2.2.2 적용 모형

문항반응이론은 등급반응모형(Samejima, 1969), 일반화 부분 점수 모형(Muraki, 1990, 1992), 명목 반응 모형(Bock, 1972)을 적용하여 추정하였다. 무응답은 0으로 코딩하여 명목 반응 모형에 대

응할 수 있도록 하였다(Thissen et al., 2010). 이 때 Bayesian 추정 방법을 이용(Chalmers, 2015)하여 집단 내 유사 응답 패턴이 반영되도록 다중 소속 다층 모형(Chung & Beretvas, 2012)을 적용하였다. 선행 연구(Walumbwa et al., 2008)가 진성리더십이 1요인임을 전제하고 있으므로 1요인으로 고정하고 추정하였다.

2.2.3 문항 타당도 및 신뢰도 검토

문항 타당도 및 신뢰도 검토는 문항 적합도 지수(item fit indices)를 검토하였다. 여러 선행 연구(Chalmers & Ng, 2017; Drasgow et al., 1985; Kang & Chen, 2008)가 제시한 문항의 가능도 계산(Likelihood calculation) 때 기대 가능도 $E(LL)$과 관찰된 LL 간 격차를 추정하여 $p<.005$ 수준(Benjamin et al., 2018)에서도 기대 값을 이탈하는 경우 문제가 있는 문항으로 파악하고 문항을 제거 후 다시 추정하는 절차를 밟았다.

2.3 최종 모형 선택

각 문항이 모형에서 추정 오차 RMSEA(Garrido, Abad, & Ponsoda, 2016; Maydeu-Olivares, 2017; Maydeu-Olivares & Joe, 2014; Maydeu-Olivares, Shi, & Rosseel, 2017)가 최소화 되는 모형을 최종 모형으로 선택하였다.

Ⅲ. 연구 결과

3.1 최종 모형

추정 결과를 Table 1, Table 2, Table 3에 제시하였다. Table 1은 16문항 중 총 5문항을 최종 모형으로 선택한 것이다. Factor Loading은 일반적으로는 절대값 기준 .3에서 .5를 넘어야 한다. 이를 제곱하면 공통분 추정치(Communality)가 되고, 1에서 공통분 추정치를 빼면 오차 추정치로 표현된다. 그러므로 Factor Loading이 1에 가까울수록 한 개념을 측정하고 있는 안정적인 문항으로 인정할 수 있다.

“나의 직속상사는 다른 사람(예: 부대원이나 타 부대 지휘관)들에게 본인의 실수에 대해 인정한다.”, “나의 직속상사는 리더로서 무엇을 해야 할 지는 도덕적 기준을 따른다.”, “나의 직속상사는 어떤 결심을 굳히기 전에 다른 사람(예: 부대원이나 타 부대 지휘관)들의 의견을 구하려고 한다.”, “나의 직속상사는 자신과 동의하지 하지 않는 사람(예: 부대원이나 타 부대 지휘관)들의 생각을 경

청한다.", "나의 직속상사는 다른 사람(예: 부대원이나 타 부대 지휘관)들의 생각을 매우 주의 깊게 경청한 후에 의사결정을 내린다." 다섯 문항은 진성 리더십을 오차 없이 측정할 수 있는 핵심 문항으로 추출되었다. 이들 문항이 왜곡 응답이나 사회적으로 바람직한 응답 등에 의해 발생할 수 있는 측정 오차에 얼마나 굳건하고 안정적인지는 문항 적합도 추정치 Table 2에 나타나있다.

Table 1. Factor Loadings and Communality estimates

Items	Factor Loadings	Communality
ALQ8: 나의 직속상사는 다른 사람(예: 부대원이나 타 부대 지휘관)들에게 본인의 실수에 대해 인정한다.	.867	.752
ALQ12: 나의 직속상사는 리더로서 무엇을 해야 할 지는 도덕적 기준을 따른다.	.871	.758
ALQ13: 나의 직속상사는 어떤 결심을 굳히기 전에 다른 사람(예: 부대원이나 타 부대 지휘관)들의 의견을 구하려고 한다.	.917	.842
ALQ14: 나의 직속상사는 자신과 동의하지 하지 않는 사람(예: 부대원이나 타 부대 지휘관)들의 생각을 경청한다.	.905	.818
ALQ16: 나의 직속상사는 다른 사람(예: 부대원이나 타 부대 지휘관)들의 생각을 매우 주의 깊게 경청한 후에 의사결정을 내린다.	.928	.861

Table 2. Item fit indices estimates

Items	Zh	S_X2	df.S_X2	RMSEA. S_X2	p.S_X2
ALQ8: 나의 직속상사는 다른 사람(예: 부대원이나 타 부대 지휘관)들에게 본인의 실수에 대해 인정한다.	1.825	29.036	16.000	.034	.024
ALQ12: 나의 직속상사는 리더로서 무엇을 해야 할 지는 도덕적 기준을 따른다.	1.733	32.404	14.000	.043	.004
ALQ13: 나의 직속상사는 어떤 결심을 굳히기 전에 다른 사람(예: 부대원이나 타 부대 지휘관)들의 의견을 구하려고 한다.	2.533	19.905	12.000	.031	.069
ALQ14: 나의 직속상사는 자신과 동의하지 하지 않는 사람(예: 부대원이나 타 부대 지휘관)들의 생각을 경청한다.	2.422	32.965	14.000	.044	.003
ALQ16: 나의 직속상사는 다른 사람(예: 부대원이나 타 부대 지휘관)들의 생각을 매우 주의 깊게 경청한 후에 의사결정을 내린다.	2.899	32.940	15.000	.041	.005

Table 2는 $E(LL)$을 표준화 하였을 때 LL이 $E(LL)$과 얼마나 멀리 떨어져 있는지를 Z값 Zh와 조정된 카이 제곱 S-X2, 조정된 카이 제곱의 자유도, 조정된 카이 제곱의 RMSEA, 조정된 카이 제곱의 p-value로 표현한 것이다. $p < .005$ 수준(Benjamin et al., 2018)을 이탈하는 경우 이상 문항으로 취급하나, 양측 검정 기준으로 Z값과 S-X2의 p-value는 $p < .0025$ 수준에서 검정 되어야 한다. 이에 대응하는 Z값은 -2.80 이하,[1] S-X2의 p-value는 .0025이다. 이 준거를 이탈하는 문항이 존재하지

않으므로 추출된 5문항이 진성 리더십을 측정하기에 적절했다고 볼 수 있다.

Table 3. Discrimination (a) and response difficulty (b_i)

Items	a (변별도)	b1 (무응답/전혀 아니다 경계면)	b2 (전혀 아니다/ 아니다 경계면)	b3 (아니다/보통 경계면)	b4 (보통/그렇다 경계면)	b5 (그렇다/매우 그렇다 경계면)
ALQ8	2.966		-3.124	-2.277	-1.117	.102
ALQ12	3.016		-2.984	-2.388	-1.157	-.035
ALQ13	3.925	-3.178	-2.671	-1.972	-1.039	.125
ALQ14	3.612	-3.213	-2.609	-1.839	-.950	.215
ALQ16	4.231		-2.600	-1.834	-.973	.046

Table 3은 검사 문항 별 변별도 a와 선택지 응답 곤란도(난이도) b_i를 표현한 것이다. 검사 문항별 난이도는 항상 양수여야 하며, 숫자가 클수록 진성 리더십의 상위 그룹과 하위 그룹을 나눌 수 있는 변별력이 크다. 또한 선택지 응답 곤란도(난이도) b_i는 MMLE(Marginal Maximum Likelihood Estimation) (Bock & Aitkin, 1981)의 선험 분포(prior distribution)인 가우시안 표준 정규분포 $\theta \sim N(0,1)$을 따르는 특성 때문에 피평가자의 절대적인 진성 리더십 위치를 나타내는 지표가 된다.

예를 들어, 평가를 받는 리더가 모든 문항에 “매우 그렇다”는 평가를 받았을 경우, 최대 $\theta = .215$에 해당하는 점수를 얻게 된다.[2] 반대로 “전혀 아니다” 또는 “무응답”이 나타났을 때는 $\theta = -3.213$에 해당하는 점수를 얻게 된다. θ를 선형 변환하여 평균이 50, 표준편차가 10인 T점수로 만들었을 때 $\mathrm{T} = 50 + \theta \times 10$의 수식으로 점수를 얻게 되므로 최소 17.87점, 최대 52.15점을 기대할 수 있다.

1) Zh 값은 양수인 경우 정상으로 취급한다(Drasgow et al., 1985, p. 77).
2) 최고점 응답 경향이 진실되지 않아 대부분의 피평가자가 높은 평가를 받았을 때에는 이를 보정하는 알고리즘이 따로 존재(Jin & Wang, 2014)한다.

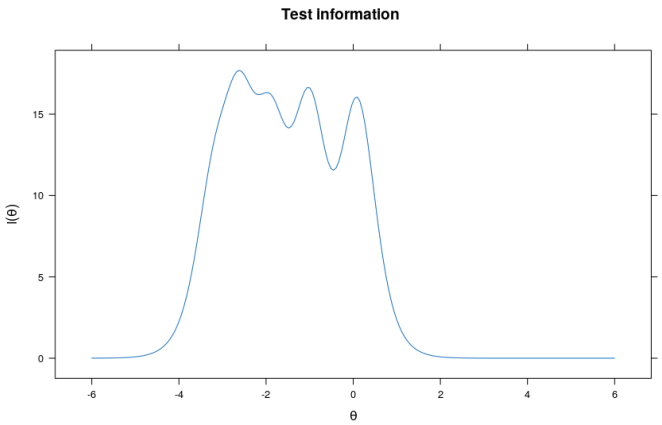

Figure 1. Range of coverage for test information

이는 +2표준편차에 해당하는 점수인 70점을 밑도는 수준이며, 진성 리더십 평가는 진성 리더십이 높은 리더를 변별하는 게 아닌, 진성 리더십이 낮은 리더를 변별하는 검사로서 기능함을 뜻한다. 이를 뒷받침하는 그림이 Figure 1이다.

Figure 1은 검사 정보 함수(TIF: Test Information Function)다. 이는 검사가 측정하고 기능할 수 있는 범위를 그림으로 나타낸 것이다. 정보 $I(\theta)$가 풍부할수록 해당 지역의 θ가 유용성이 있음을 보여준다. 그런데 이 그림에서 보듯 $-4 < \theta < 1$ 정도의 범위에서 진성 리더십 검사가 동작하고 있음을 알 수 있다. 이는 진성 리더십 검사가 진성 리더십이 낮은 리더를 변별하는 데 유용한 검사로 사용됨을 증명하는 것이다. 또한 마찬가지의 맥락으로 Figure 2는 진성 리더십 검사가 어떤 범위에서 신뢰도가 확보될 수 있는지 나타내고 있다. 신뢰도 또한 $-4 < \theta < 1$ 정도 범위에서 .6 이상이 확보되고 있으며, 더 높은 점수를 얻는 경우 또는 더 낮은 점수를 얻는 경우 신뢰성이 떨어짐을 알 수 있다. 따라서 이는 진성 리더십 검사가 낮은 진성 리더십 수준을 지닌 리더를 판별하는 데 유용함을 다시 증명한다.

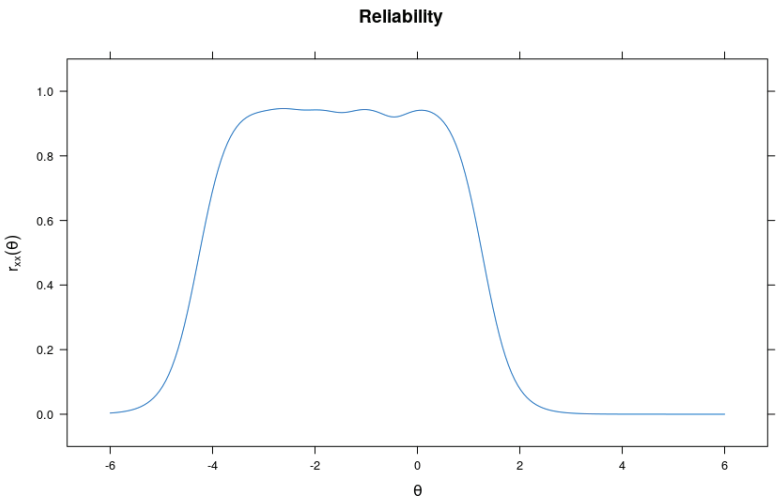

Figure 2. Reliability Function $\boldsymbol{r_{xx}(\theta)}$ and coverage area of $\boldsymbol{\theta}$

그렇다면 왜 진성 리더십 검사는 “낮은” 수준의 진성 리더십을 지닌 리더를 변별하는 것일까?

첫번째 가능성은 5점 응답 척도(Likert, 1932)가 원 검사 개발 연구(Walumbwa et al., 2008)에서 서열 척도 또는 명목 척도로 고려되지 않았기 때문에 알 수 없었을 가능성이 높다. 최근 구조방정식모형 연구에서는 2점부터 7점까지 Likert 척도를 토대로 시뮬레이션 연구를 진행한 바 있다(Rhemtulla, Brosseau-Liard, & Savalei, 2012b). 이 연구의 결론은 2점에서 5점 척도까지는 연속 척도가 아닌 서열 척도로 취급해야 하며, 6점이 넘을 때서야 연속 척도로 고려해야 하는 점이다. 구조방정식모형에서 응답을 서열 척도로 취급하는 건 문항반응이론 사용을 뜻한다(Wirth & Edwards, 2007).

두 번째 가능성은 5점 선택지 자체가 주는 모호함이 있었을 것이다. Wakita와 동료(Wakita, Ueshima, & Noguchi, 2012)는 4점, 5점, 7점 선택지를 주고 선택지 개수 별 정확성을 추정하였다. 그 결과 4점이 가장 정확한 측정이 가능했음을 보였고, 선택지 개수가 늘어날수록 부정확한 추정 결과를 보여주었다.

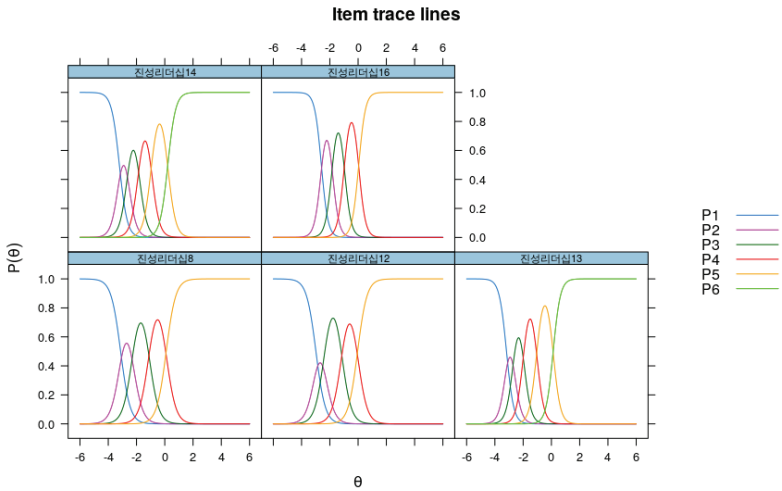

Figure 3. Item trace lines of Authentic Leadership Items

이를 뒷받침하듯 Figure 3은 Trace line의 P1과 P2, 또는 P2와 P3가 지나치게 가깝게 배치돼 있다. 특히 ALQ12, ALQ13은 P2가 차지하는 고유 면적이 지나치게 좁아, ALQ12는 '아니다', ALQ13은 '전혀 아니다'가 사실상 필요하지 않은 선택지였음을 보이고 있다. 또한 5점에 해당하는 '매우 그렇다'는 검사 문항이 평범하여 선택이 쉬웠을 것이다. 문항반응이론의 논리 중 하나는 누적응답기제(cumulative response process)이다(Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). 누적응답기제에서 최고점(Full credit)을 받는 사람이 많으면 최고점의 가치는 감쇠(attenuation)된다. 따라서 Figure 3에서 보듯 P5 (또는 P6)가 $\theta = 0$ 언저리에 위치해 있는 것이다.

Ⅳ. 결론 및 제언

목숨을 잃을 수도 있고, 극한의 거주 환경 속에 놓인 군 현장에서 적응하기 위해서라도 "진정성 있는 리더"는 언제나 의미 있는 존재(Gaddy, Gonzalez, Lathan, & Graham, 2017)이며, 쓰러져도 다시 일어나는 힘을 준다. 이는 반대로 말하면 진정성 없는 리더는 군 현장에서 큰 도움이 되지 않으며, 고통스러운 상황에서 부하의 적응과 회복을 돕는 데 도움이 되진 않을 수 있음(Gaddy et al., 2017)을 내포할 수도 있다. Gaddy와 동료(2017)가 지적한 것처럼 진정성 없는 리더는 "사이코패시"(psychopathy)가 높은 사람일 수 있으며, 이런 "사이코패시"가 높은 리더를 최소화 하기 위해서라도 진성 리더십은 중요하다(Gaddy et al., 2017).

다만 진성 리더십이 높은 사람을 변별할 수 있는지, 아니면 진성 리더십이 낮은 사람을 변별할 수 있는지조차 알려지지 않았던 상황에서 이 연구는 진성 리더십 검사가 진성 리더십이 낮은 사람

을 변별해 내는 기능이 있을 수 있음을 보여주었다. 검사로 진성 리더십이 높은 리더에게 상을 줄 수는 없지만, 진성 리더십이 극단적으로 낮은 사람(T점수 30점 이하)은 승진하지 못하게 하거나 리더로서 재심을 하게 함으로서 병사의 사기를 저해하지 못하도록 해야 한다(Gaddy et al., 2017).

향후 연구는 쌍요인(Bifactor) 해(Gibbons & Hedeker, 1992; Mansolf & Reise, 2016)를 검토함으로서 이 연구의 정밀도를 높이고, 더 많은 시사점을 구축할 필요가 있다. 이 연구를 시발점으로 군 장면에서 진정성 리더십을 발휘하지 못하는 리더와 진정성 리더십을 발휘하는 리더를 구별하고 인력 운용에 어떻게 반영할 수 있을지 고민을 지속할 필요가 있겠다.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove, Illinois: Waveland.

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125 – 141. https://doi.org/10.1177/0265532212452396

Bae, S. (2017). *kaefa: kwangwoon automated exploratory factor analysis*. Retrieved from https://github.com/seonghobae/kaefa

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6 – 10. https://doi.org/10.1038/s41562-017-0189-z

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29 – 51. https://doi.org/10.1007/BF02291411

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*(4), 21 – 33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443 – 459. https://doi.org/10.1007/BF02293801

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing Rater Biases in 360-Degree Feedback by Forcing Choice. *Organizational Research Methods*, *20*(1), 121 – 148. https://doi.org/10.1177/1094428116668036

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 309 – 329. https://doi.org/10.1348/000711007X249603

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, *57*(1), 15 – 22. https://doi.org/10.1037/h0034185

Cao, M., Drasgow, F., & Cho, S. (2015). Developing Ideal Intermediate Personality Items for the Ideal Point Model. *Organizational Research Methods*, *18*(2), 252 – 275. https://doi.org/10.1177/1094428114555993

Carnes, A., Houghton, J. D., & Ellison, C. N. (2015). What matters most in leader selection? The role of personality and implicit leadership theories. *Leadership & Organization Development Journal*, *36*(4), 360 - 379. https://doi.org/10.1108/LODJ-06-2013-0087

Carter, N. T., Dalal, D. K., Boyce, A. S., O'Connell, M. S., Kung, M.-C., & Delgado, K. M. (2014). Uncovering curvilinear relationships between conscientiousness and job performance: How theoretically appropriate measurement makes an empirical difference. *Journal of Applied Psychology*, *99*(4), 564 - 586. https://doi.org/10.1037/a0034688

Chalmers, R. P. (2015). Extended Mixed-Effects Item Response Models with the MH-RM algorithm. *Journal of Educational Measurement*, *52*(2), 200 - 222. https://doi.org/10.1111/jedm.12072

Chalmers, R. P., & Ng, V. (2017). Plausible-Value Imputation Statistics for Detecting Item Misfit. *Applied Psychological Measurement*, *41*(5), 372 - 387. https://doi.org/10.1177/0146621617692079

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting Item Response Theory Models to Two Personality Inventories: Issues and Insights. *Multivariate Behavioral Research*, *36*(4), 523 - 562. https://doi.org/10.1207/S15327906MBR3604_03

Chung, H., & Beretvas, S. N. (2012). The impact of ignoring multiple membership data structures in multilevel models. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 185 - 200. https://doi.org/10.1111/j.2044-8317.2011.02023.x

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67 - 86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Ferrando, P. J. (2009). Multidimensional Factor-Analysis-Based Procedures for Assessing Scalability in Personality Measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(1), 109 - 133. https://doi.org/10.1080/10705510802561352

Gaddy, J. W., Gonzalez, S. P., Lathan, C. A., & Graham, P. K. (2017). The Perception of Authentic Leadership on Subordinate Resilience. *Military Behavioral Health*, *5*(1), 64 - 72. https://doi.org/10.1080/21635781.2016.1243495

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are Fit Indices Really Fit to Estimate the Number of Factors With Categorical Variables ? Some Cautionary Findings via Monte Carlo Simulation. *Psychological Methods*. https://doi.org/10.1037/met0000064

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*,

57(3), 423 - 436. https://doi.org/10.1007/BF02295430

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review*, *36*(3), 341 - 355. https://doi.org/10.1108/00483480710731310

Henson, R. K., & Roberts, J. K. (2006). Use of Exploratory Factor Analysis in Published Research. *Educational and Psychological Measurement*, *66*(3), 393 - 416. https://doi.org/10.1177/0013164405282485

Holden, R. R., & Book, A. S. (2009). Using hybrid Rasch-latent class modeling to improve the detection of fakers on a personality inventory. *Personality and Individual Differences*, *47*(3), 185 - 190. https://doi.org/10.1016/j.paid.2009.02.024

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1 - 55. https://doi.org/10.1080/10705519909540118

Ilies, R., Morgeson, F. P., & Nahrgang, J. D. (2005). Authentic leadership and eudaemonic well-being: Understanding leader - follower outcomes. *The Leadership Quarterly*, *16*(3), 373 - 394. https://doi.org/10.1016/j.leaqua.2005.03.002

Jin, K.-Y., & Wang, W.-C. (2014). Generalized IRT Models for Extreme Response Style. *Educational and Psychological Measurement*, *74*(1), 116 - 138. https://doi.org/10.1177/0013164413498876

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X^2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391 - 406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Kang, T., & Chen, T. T. (2011). Performance of the generalized S-X2 item fit index for the graded response model. *Asia Pacific Education Review*, *12*(1), 89 - 96. https://doi.org/10.1007/s12564-010-9082-4

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The Performance of RMSEA in Models With Small Degrees of Freedom. *Sociological Methods & Research*, *44*(3), 486 - 507. https://doi.org/10.1177/0049124114543236

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford.

Kwak, J., & Choi, K. (2016). A Study on the Direction of the Selection Tools for Entry level Leaders of the Army. *KIDA Defense Issues & Analyses*, *16*(40), 1 - 11. Retrieved from www.kida.re.kr/frt/board/frtNormalBoardDetail.do?sidx=382&idx=1690&depth=4

LaHuis, D. M., Clark, P., & O'Brien, E. (2011). An Examination of Item Response Theory Item

Fit Indices for the Graded Response Model. *Organizational Research Methods*, *14*(1), 10 - 23. https://doi.org/10.1177/1094428109350930

Lai, K., & Green, S. B. (2016). The Problem with Having Two Watches: Assessment of Fit When RMSEA and CFI Disagree. *Multivariate Behavioral Research*, *51*(2 - 3), 220 - 239. https://doi.org/10.1080/00273171.2015.1134306

Lee, D., & Park, Y. (2017). A Study on the Improvement of Interview System for the Selection of Entry level leaders. *KIDA Defense Issues & Analyses*, *17*(29), 1 - 8. Retrieved from http://www.kida.re.kr/frt/board/frtNormalBoardDetail.do?sidx=382&idx=1735&depth=4

Levashina, J., Weekley, J. A., Roulin, N., & Hauck, E. (2014). Using Blatant Extreme Responding for Detecting Faking in High-stakes Selection: Construct validity, relationship with general mental ability, and subgroup differences. *International Journal of Selection and Assessment*, *22*(4), 371 - 383. https://doi.org/10.1111/ijsa.12084

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 1 - 55. Retrieved from https://legacy.voteview.com/pdf/Likert_1932.pdf

Liu, C.-W., & Chalmers, R. P. (2018). Fitting item response unfolding models to Likert-scale data using mirt in R. *PLOS ONE*, *13*(5), e0196292. https://doi.org/10.1371/journal.pone.0196292

Makransky, G., & Glas, C. A. W. (2013). Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement*, *46*(9), 3228 - 3237. https://doi.org/10.1016/j.measurement.2013.06.020

Makransky, G., Mortensen, E. L., & Glas, C. A. W. (2013). Improving Personality Facet Scores With Multidimensional Computer Adaptive Testing. *Assessment*, *20*(1), 3 - 13. https://doi.org/10.1177/1073191112437756

Mansolf, M., & Reise, S. P. (2016). Exploratory Bifactor Analysis: The Schmid-Leiman Orthogonalization and Jennrich-Bentler Analytic Rotations. *Multivariate Behavioral Research*, *51*(5), 698 - 717. https://doi.org/10.1080/00273171.2016.1215898

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*. https://doi.org/10.1037/a0019227

Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*, *49*(6), 1194 - 1218. https://doi.org/10.1037/a0026913

Maydeu-Olivares, A. (2017). Assessing the Size of Model Misfit in Structural Equation Models. *Psychometrika*, 1 - 26. https://doi.org/10.1007/s11336-016-9552-7

Maydeu-Olivares, A., & Joe, H. (2014). Assessing Approximate Fit in Categorical Data Analysis.

Multivariate Behavioral Research, *49*(4), 305 - 328. https://doi.org/10.1080/00273171.2014.911075

Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2017). Assessing Fit in Structural Equation Models: A Monte-Carlo Evaluation of RMSEA Versus SRMR Confidence Intervals and Tests of Close Fit. *Structural Equation Modeling*. https://doi.org/10.1080/10705511.2017.1389611

Morin, A. J. S., Katrin Arens, A., & Marsh, H. W. (2015). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, *23*(1). https://doi.org/10.1080/10705511.2014.961800

Muraki, E. (1990). Fitting a Polytomous Item Response Model to Likert-Type Data. *Applied Psychological Measurement*, *14*(1), 59 - 71. https://doi.org/10.1177/014662169001400106

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i - 30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879 - 903. https://doi.org/10.1037/0021-9010.88.5.879

Podsakoff, P. M., & Organ, D. W. (1986). Self-Reports in Organizational Research: Problems and Prospects. *Journal of Management*, *12*(4), 531 - 544. https://doi.org/10.1177/014920638601200408

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*(1), 28 - 56. https://doi.org/10.1080/00273171.2012.710386

Reise, S. P., & Flannery, P. (1996). Assessing Person-Fit on Measures of Typical Performance. *Applied Measurement in Education*, *9*(1), 9 - 26. https://doi.org/10.1207/s15324818ame0901_3

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012a). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354 - 373. https://doi.org/10.1037/a0029315

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012b). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354 - 373. https://doi.org/10.1037/a0029315

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

Psychometrika, *34*(S1), 1 - 97. https://doi.org/10.1007/BF03372160

Sinharay, S. (2015). Assessment of Person Fit for Mixed-Format Tests. *Journal of Educational and Behavioral Statistics*, *40*(4), 343 - 365. https://doi.org/10.3102/1076998615589128

Tay, L., & Drasgow, F. (2012). Adjusting the Adjusted χ 2 / df Ratio Statistic for Dichotomous Item Response Theory Analyses. *Educational and Psychological Measurement*, *72*(3), 510 - 528. https://doi.org/10.1177/0013164411416976

Tendeiro, J. N. (2017). The lz(p)* Person-Fit Statistic in an Unfolding Model Context. *Applied Psychological Measurement*, *41*(1), 44 - 59. https://doi.org/10.1177/0146621616669336

Thissen, D., Cai, L., & Bock, R. D. (2010). The Nominal Categories Item Response Model. In M. L. Nering & R. Ostini (Eds.), *Handbook of Polytomous Item Response Theory Models*. https://doi.org/10.4324/9780203861264.ch3

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological Distance Between Categories in the Likert Scale: Comparing Different Numbers of Options. *Educational and Psychological Measurement*, *72*(4), 533 - 546. https://doi.org/10.1177/0013164411431162

Walumbwa, F. O., Avolio, B. J., Gardner, W. L., Wernsing, T. S., & Peterson, S. J. (2008). Authentic Leadership: Development and Validation of a Theory-Based Measure. *Journal of Management*, *34*(1), 89 - 126. https://doi.org/10.1177/0149206307308913

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58 - 79. https://doi.org/10.1037/1082-989X.12.1.58

Yang, K. H., Kim, J., & Kim, J. (2017). Effects of authentic leadership on organizational commitment. *Journal of Human Resource Management Research*, *24*(2), 1 - 26. https://doi.org/10.14396/jhrmr.2017.24.2.1

원 고 접 수 일 2019년 4월 6일
원 고 수 정 일 2019년 4월 25일
게 재 확 정 일 2019년 5월 8일

Abstract

https://doi.org/10.37944/jams.v2i1.41

Which Item Measure the Authentic Leadership of the Entry Level Leaders in Military Appropriately?

Seongho Bae

Kwangwoon University

Due to little research focusing on high precise measurement of authentic leadership is not active, and haven't knowledge for reliability and coverage range of authentic leadership questionnaires when want to make high stakes personnel decisions in military settings. This study presented five questionnaires to measure authentic leadership appropriately based on military leader responded survey data. They were ALQ8, ALQ12, ALQ13, ALQ14, and ALQ16. In this study presented they could be discriminate leaders who low level of authentic leadership with reliability. Further tasks may discover and consider of bi-factor solution of factor structure of authentic leadership questionnaire, and flourish discussions how to use authentic leadership questionnaire to make decisions in context of military leader resource management.

Keywords : IRT, Appropriate Measurement, Coverage, Item Selection