고정항목을 활용한 국가 간 경찰신뢰 측정 연구*

**

차 례

〈국 문 초 록〉

이 연구는 King et al.(2004)이 제기한 문제 인식을 바탕으로, 국가 간 비교 연구에서 발생하는 측정 오류인 차별문항작용(Differential Item Functioning) 문제를 해결하기 위해, 고정항목을 활용한 경찰 신뢰 측정의 보정 방안을 제시하고 그 타당성을 실증적으로 검증하는 것을 목적으로 한다. 이 연구는 세계가치조사(WVS) 국제설문 프로젝트(1981-2022) 데이터를 활용하여, 설문 내 신뢰 항목(가족, 이웃, 타인)을 고정항목으로 활용하는 비모수적 보정 방법론을 제안하였다. 분석은 전 세계 95개국을 대상으로 수행되었으며, 비모수적 부트스트랩 기법을 적용하여 신뢰구간을 추정하였다. 분석 결과, 영국과 일본의 사례에서 확인되듯 원점수 기반의 비교는 국가별 응답 성향에 의해 신뢰 수준이 역전되거나 왜곡될 수 있음을 확인하였다. 특히 핀란드, 니카라과, 네덜란드 등 주요 변동 국가들의 보정 후 순위 변화는 단순 표본오차를 넘어서는 통계적 유의성을 보였다. 이 연구는 정책학 및 행정학 분야의 국제 비교 연구에서 차별문항작용 문제에 대한 통제 필요성을 환기하고, 고정항목을 활용한 보정 절차와 그 효용 및 한계를 제시했다는 점에서 학술적·정책적 의의를 지닌다.

주제어: 차별문항작용, 경찰 신뢰, 국가간 비교, 고정항목, World Value Survey

Ⅰ. 서 론

경찰이 국가기관으로써 권위와 정당성을 유지하기 위해서는 시민들의 신뢰가 요구된다(Bittner, 1970), 경찰이 신뢰를 잃게 되면 시민의 생명과 재산, 신체를 보호하고 사회의 안녕과 질서를 유지하는 역할을 효율적이고 효과적으로 수행하는데 지장을 받게 된다(Hamm et al., 2017; Jackson & Bradford, 2010; Sunshine & Tyler, 2003). 지난 40년 동안 국제간 설문조사 프로젝트(World Value Survey, Global Barometer, Pew Research Center 등)는 경찰에 대한 신뢰라는 질문을 전세계 국가를 대상으로 제기하고, 설문조사하여 자료를 축적하고 있다.

많은 연구자들은 국제간 설문조사 프로젝트에서 축적한 데이터를 활용하여 국가별 또는 연도별 경찰신뢰 질문항의 평균치를 구하여 그 국가의 경찰신뢰 수준으로 측정하고, 이를 다른 국가와 비교하는 연구를 진행하였다(Cao, Lai, & Zhao, 2012; Jang et al., 2010; Cao & Zhao, 2005; Schaap & Scheepers, 2014; 장현석, 2014; 정보성・이창배, 2018). 이러한 선행 연구들은 대체로 국가마다 경찰신뢰에 차이가 있고 역동적으로 변화한다는 점을 공유하고 있다(Schaap & Scheepers, 2014).

하지만, 집단 간의 단순 평균을 비교하는 전통적 측정방법으로 국가 간 또는 집단 간의 속성과 태도를 비교 연구하는 것에 의문을 제기하는 학자들도 존재한다. King et al. (2004), Wand (2013), Aldrich & McKelvey(1977)이 대표적인데, 대인간의 비교불가능성으로 요약되는 차별문항작용(Differential Item Functioning, 이하 DIF)의 문제이다. 예컨대, 경찰신뢰를 측정하려고 할 때, 어떤 국가의 응답자들은 기본적으로 신뢰하는 경향이 높고, 어떤 국가의 응답자들은 기본적으로 신뢰하는 경향이 낮다는 점을 인정한다면, 경찰신뢰를 측정하는 서열척도의 각 기준점의 위치가 다를 수 있다. 그렇다고 한다면, 두 국가의 평균 경찰신뢰 수준이 동일하다고 하더라도, 해당 국가의 신뢰 경향에 따라 다르게 해석해야 할 여지가 있을 것이다.

이 연구는 King et al. (2004)이 제기한 차별문항작용의 문제 인식을 바탕으로, 국가간 비교 연구에서 응답자의 속성과 태도 측정에 대한 전통적 측정방법의 대안으로 고정항목을 활용한 비모수적 측정 방법을 제안하는 것을 목적으로 한다. 이는 단순히 설문 데이터의 수치를 비교하는 수준을 넘어, 각 국가의 문화적 맥락이 응답 척도에 미치는 영향을 통계적으로 보정함으로써 비교 연구의 과학적 타당성을 높이려는 시도이다. 이를 위해 World Value Survey 국제설문 프로젝트의 데이터를 활용하여 전 세계 95개 국가의 경찰 신뢰도를 분석하였다. 실증 분석 결과, 원자료를 그대로 활용하는 전통적인

측정방법과 고정항목을 활용한 새로운 측정방법 간에는 전반적으로 높은 상관관계가 형성됨을 확인하였다. 그러나 응답자의 주관적 기준이나 문화적 배경이 상이한 다수의 국가에서는 두 방식에 따른 신뢰도 수치와 순위에서 유의미한 차이를 발견되었다. 이러한 분석 결과는 설문 원자료를 단순 비교하는 방식이 국가 간 비교의 타당성을 저해할 수 있음을 보여주며, 고정항목을 활용한 보정 방식이 이를 효과적으로 보완할 수 있음을 의미한다. 결론적으로 본 연구는 제안된 방법론이 국제 비교 연구의 정밀도를 높이고 새로운 분석적 통찰을 제공할 수 있을 것으로 기대한다.

Ⅱ. 차별문항작용(DIF)의 문제

1. 개인 태도 측정의 한계

사회과학자는 설문조사를 통해 개인의 태도와 속성을 측정하고 한다. 하지만, 설문조사에는 몸무게를 측정하는 '저울'이나 키를 측정하는 '자'와 같은 엄격한 의미의 정량화된 척도는 없다. 그럼에도 불구하고, 응답자 개인은 연구자가 관심을 갖는 개념에 대한 태도나 속성을 가지고 있기 때문에, 사회과학자는 그 태도와 속성을 측정하고자 노력한다. 예컨대, 행정기관에 대한 신뢰의 수준을 파악하고자 한다면, 설문지를 작성하여 응답자 개인에게 스스로 평가할 수 있도록 4점 척도(매우 불신, 불신, 신뢰, 매우 신뢰) 또는 5점 척도(매우 불신, 불신, 중립, 신뢰, 매우 신뢰)와 같은 일련의 순서가 있는 범주를 제공한다. 이를 통해 연구자는 응답자의 태도와 속성을 간접적으로 관찰하게 된다.

그런데, 연구자는 순서형 범주를 통해 개인 수준 데이터를 분석하는 과정에서 발생할 수 있는 방법론적 한계에 봉착하게 된다. 신뢰에 대한 '4점 척도'에서 발생할 수 있는 데이터 유형의 예를 살펴보자. 〈표 1〉의 예는 두 명의 가상 응답자가 가질 수 있는 인식을 보여준다. 이 가상의 두 응답자는 신뢰 척도에서 신뢰 대상(여기서는 처음 만난 사람, 이웃, 가족)의 위치를 정하는 데 있어 큰 차이를 보이지만, 신뢰 대상자가 속한 근본적인 척도의 순서는 상당히 일치하는 것으로 보인다. 구체적으로 살펴보면, 응답자 1은 가족에 대해서는 "매우 신뢰한다", 이웃에 대해서는 "신뢰한다", 처음 만난 사람에 대해서는 "신뢰하지 않는다" 순으로 신뢰하고, 응답자 2는 가족에 대해서는 "신뢰한다", 이웃에 대해서는 "신뢰하지 않는다", 처음 만난 사람에 대해서는 "매우 신뢰하지 않는다" 순으로 신뢰라는 척도에 위치시킨다.

<표 1> 가상적인 응답자의 4점 신뢰 척도

	"불신"	1 (매우 불신)	2 (불신)	3 (신뢰)	4 (매우 신뢰)	"신뢰"
응답자 1			P, S	N	F	
응답자2		P	N, S	F		

주: F = Family(가족에 대한 신뢰), N = Neighbor(이웃에 대한 신뢰), P = People you meet for the first time(처음 만난 사람에 대한 신뢰), S = Self(경찰이라는 신뢰 대상에 대한 자기 응답)

이때 응답자들이 신뢰 대상의 척도상 위치를 정하는 데 있어 의견이 일치하지 않을 수 있다. 그 이유는 응답자들은 신뢰 정도에 대한 각자의 이해와 해석에 따라 척도의 기준점을 다르게 설정할 뿐만 아니라, 척도의 구간을 다르게 해석할 수 있기 때문이다. 다시 말해, 응답자들은 신뢰대상에 대한 인식을 통한 자신의 신뢰척도를 "신뢰" 범주(예: 3, 4)에 더 자주 배치하거나, "비신뢰" 범주(예: 1, 2)에 더 자주 배치하는 경향이 발생할 수 있다.

이러한 경향은 데이터 분석 시 특이한 결과를 초래한다. 응답자 1과 응답자 2는 모두 연구자가 관심갖는 변수에 대한 응답(<표 1>의 경우 경찰에 대한 신뢰)이 모두 2번째 범주("신뢰하지 않는다")를 답하였다. 그런데, 응답 2의 의미가 응답자 1과 응답자 2에게 확연히 다를 수 있다. 그것은 "2"라는 숫자가 등간척도 또는 비율척도로 측정된 것이 아니라 서열척도로 측정되었고, 그 범주 간의 간격이 응답자에 따라 상당히 다를 수 있기 때문이다. 이 경우에 응답자 1과 응답자 2가 경찰에 대한 신뢰에 부여한 "2"라는 숫자를 어떻게 이해하고 해석하고 조정할 것인가 하는 것이 이 연구의 핵심적 주제이다.

2. 응답자의 서열척도 관찰 매커니즘

개인의 태도와 속성을 설문조사를 통해 관찰하는 과정에서, 응답자 간 다른 인식과 이해로 인해 차별화된 서열척도가 어떻게 관측치로 나타나는지를 설명하고자 한다. 여기서 서열척도는 순서는 뚜렷하지만 그 순서 간의 간격은 모호하여 관측되지 않아서 개인의 마음속 근저에 남아있다는 점에서 잠재적이라고 할 수 있다. 즉, 1로 표현된 "매우 신뢰하지 않는다", 2로 표현된 "신뢰하지 않는다", 3으로 표현된 "신뢰한다", 4로 표현된 "매우 신뢰한다"에서 1, 2, 3, 4라는 '눈금'의 위치가 개인별로 차이가 나지만 이는 관측되지 않고 응답자 개인의 마음속에 놓여 있다는 점에서 1, 2, 3, 4의 위치는 '잠재적'이다.

여기서 응답범주인 "신뢰한다"거나 "신뢰하지 않는다"를 설문지에서 명확하게 정의하지 않았기 때문에, 그리고 설령 설명하였다고 하더라고 응답자가 다르게 해석하고 받아들일 수 있기 때문에, 각각의 응답범주는 다양한 상황을 포괄할 수 있다. 심지어 "매우 신뢰하지 않는다"라는 범주조차도 개인의 기준에 따라 기준점이 다르게 정해질 수 있다. 통계 모형은 일반적으로 서열 척도를 1차원 실수선의 이산적 표현으로 모형화하게 되는데, 범주는 실수선를 나누는 상호배타적인 간격 집합으로 정의된다. 잠재적인 서열척도를 나타내는 근저의 실수선은 연구자가 실수선 상의 개인의 위치를 관찰하는 것이 아니라, 설문조사 응답의 범주만 관찰하기 때문에 일반적으로 '잠재' 척도라고 부를 수 있다. 응답자는 신뢰라고 하는 잠재척도를 가지고 '경찰'이라는 대상에 투영하여 '경찰신뢰'의 정도를 응답하게 되며, 연구자는 응답자의 경찰신뢰 점수(1, 2, 3, 4)를 관찰할 수 있지만 응답자의 잠재척도를 관찰할 수는 없다.

〈그림 1〉는 연속적인 잠재척도에서 정의된 값을 서수 범주로 위치시키는 논리를 보여준다. 응답자(i)의 속성은 잠재척도 실수선($-\infty \sim \infty$)에서 한 지점($\tilde{y}_i$)에 위치한다. 응답자는 자신의 값 $\tilde{y}_i$를 알고 있으나, 이 값이 얼마인지를 설명할 수 없다. 다만, 응답자가 응답함으로써 관찰될 경우 이 값에 따라 순서가 정해질 뿐이다. 결국, 응답자가 서수 척도에 응답할 때 먼저 잠재 척도를 구간으로 나누는 구분점(τ)의 위치를 선택함으로써 범주의 의미를 정의해야 한다. 범주 k와 $k+1$을 구분하는 구분점은 τ_{ik}로 표기한다. 따라서 관찰된 범주의 선택(y_i)은 다음과 같이 정의할 수 있다.

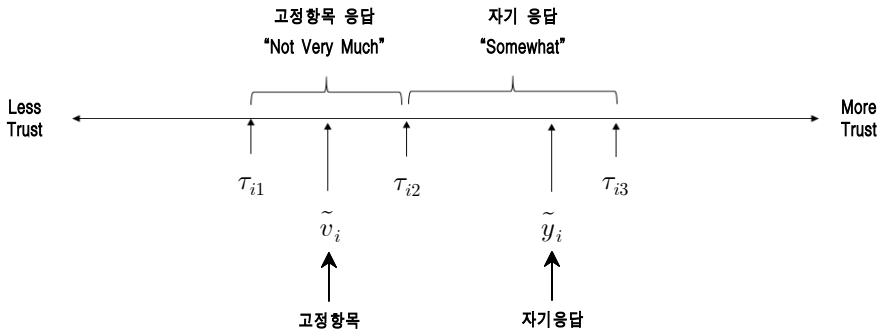
$$y_i = k \;\; if \; \tau_{i,k-1} \leq \tilde{y}_i < \tau_{ik} \qquad (1)$$

여기서 y_i를 관찰된 "자기 응답"이라고 부른다. 결국, 〈그림 2〉와 같이, $\tilde{y}_i$에 위치한 속성을 가진 응답자는 자신의 신뢰를 잠재척도 실수선 상에서 "다소 신뢰한다(somewhat)"고 응답할 것이다. 이러한 메커니즘을 통해 연구자는 응답자가 응답한 "다소 신뢰한다"를 관찰하게 되는 것이다. 그런데, 연구자는 응답자의 잠재척도 상의 구분점(τ_{ik})를 관찰하지 못한다는 점에서 차별문항작용(DIF)의 문제가 발생할 수 있다. 이 연구에서는 응답자가 설문지의 다른 문항들에서 응답한 "신뢰"의 기준점(τ_{ik})을 바탕으로 응답자의 잠재척도를 추정한다. 여기서 자기응답 항목의 기준점(τ_{ik})을 확인하기 위하여 사용되는 같은 설문지의 다른 문항을 "고정항목(anchoring items)"이라고 정의한다. 고정항목을 활용한 잠재척도(τ)의 추정, 그리고 이를 통한 자기 응답의 조정 방법에

대해서는 제3장에서 자세히 다루고자 한다.

우선, 고정항목으로 활용되기 위해서는 관심있는 자기 응답 척도와 동일한 척도를 사용하여야 한다. 즉, 자기 응답 척도가 4점 척도(예컨대, 매우 불신, 불신, 신뢰, 매우 신뢰)라고 한다면, 고정항목의 척도도 4점 척도이어야 한다. 척도의 문구가 완벽히 동일하면 좋지만, 언어적인 차이로 인해 또는 질문의 차별성으로 인해 다소 다르다고 하더라도 전체적으로 맥락에서 유사하다면 문제없다. 예컨대, 관심있는 자기 응답 척도가 "매우 불신", "불신", "신뢰", "매우 신뢰"이고, 고려 가능한 고정항목의 척도가 "매우 믿지 않는다", "다소 믿지 않는다", "어느 정도 믿는다", "매우 믿는다"라고 하더라도 고정항목으로 활용하기에 충분하다.

<그림 1> "신뢰" 잠재 척도를 서열 범주로 관측하는 메커니즘



주: Wand(2012: 251) Figure 2 원용

고정항목은 자기 응답에 사용된 순서 범주의 정의와 동일한 잠재 척도를 사용하여 평가되도록 사전에 설계될 수도 있고, 사후에 발견될 수도 있다. 고정항목(m)에 대한 관찰 평점 v_{im}은 자기 응답과 동일한 방식으로 정의된다.

$$v_{im} = k \;\; if \;\; \tau_{i,k-1} \leq \tilde{v}_{im} < \tau_{ik} \qquad (2)$$

여기서 $\tilde{v}_{im}$은 응답자(i)가 실수선 잠재척도 상에서 고정항목 m의 위치에 대해 내린 판단이다. 예컨대, 〈그림 2〉에서 응답자(i)는 $\tilde{v}_i$에 위치한 고정항목을 "별로 신뢰하지 않는다(Not very much)"로 평가한 것이다.

서열 척도로 측정된 자기 응답을 비교하는 표준적인 접근 방식은 응답자들이 각 기준

점을 어디에 두어야 하는지에 대해 공통된 이해를 공유한다고 가정한다 ($\tau_k = \tau_{ik} = \tau_{i'k}$). 그러나 주관적이거나 모호하게 정의된 서열 척도의 맥락에서 개인 간 척도 정의의 동질성 가정은 적어도 논쟁의 여지가 있다. 신뢰 경향이 강한 집단 응답자(<표 1>의 가상 응답자1)의 "신뢰하지 않는다"는 개념이 신뢰 경향이 약한 다른 집단 응답자(<표 1>의 가상 응답자2)의 "신뢰하지 않는다"라는 개념과 합리적으로 일치하지 않을 수 있다. 이것이 다음에서 설명하게 될 차별문항작용의 문제이다.

3. 차별문항작용(DIF)의 문제와 그 해결노력

응답자의 잠재척도 관찰 매커니즘을 반영하여 경찰신뢰에 대한 데이터 생성과정을 정확하게 이해하려면, 경찰신뢰에 대한 인식이 완전히 일치하더라도 잠재척도나 고정점(τ_{ik})에 대한 해석이 다르기 때문에 경찰신뢰에 대한 응답의 일치가 이루어지지 않거나 잘못 평가할 수 있다는 점을 이해하여야 한다. King et al.(2004)은 이러한 유형의 문제를 차별문항작용(DIF)라는 통계적 개념으로 정의하였다. 다시 말해, 잠재척도의 어떤 기준점의 위치가 개인마다 다를 경우($\tau_{ik} \neq \tau_{i'k}$), i와 i'의 응답은 차별문항작용(differential item functioning; King et al., 2004)의 영향을 받게 된다. 차별문항작용은 대인간 비교 불가능성을 말하는데, 응답자가 설문문항에 대한 척도를 다르게 해석하고 응답할 때 발생한다. 같은 의견을 가진 두 응답자라도 잠재척도에서 서로 다른 위치에 자신의 응답을 위치시킬 수 있으며, 반대로 서로 다른 의견을 가진 두 응답자라도 같은 위치에 자신의 응답을 위치시킬 수 있다. 이 문제는 주로 방법론적인 문제처럼 보일 수 있지만, 중요한 실질적 함의를 지닌다. 특히, 인식에 대한 국가 간 비교연구에서 국가별로 문화와 제도의 차이로 인해 발생하는 잠재척도에 대한 차별적 이해는 국가 간 응답자 간의 비교 가능성을 제한하고 시민들이 경찰에 대해 갖는 신뢰 인식의 수준을 정확히 이해하지 못하게 방해한다.

실제로 경찰에 대한 신뢰 평가의 모든 변동성을 '차별문항작용'으로 설명할 수는 없다. 하지만, 실제 경찰에 대한 신뢰($\tilde{y}_i$)의 변동으로 인한 신뢰평가(y_i)의 변동과 차별문항작용으로 인한 변동을 구분할 수 있다면 경찰신뢰에 대한 진실에 더욱 가깝게 접근할 수 있을 것이다. 실제로 DIF 문제를 해결하기 위한 노력은 크게 Aldrich & McKelvy(1977)에 개발된 A-M 방법과 King et al.(2004)에 의해 개발된 Anchoring Vinettes 방법이 있다. A-M방법은 원자료의 배치값을 자극의 진정한 위치에 대한 선형 왜곡으로 처리하여 각 응답자의 지각 왜곡 매개변수를 추정하여 응답자의 이상점을 찾

는 방법이고, Anchoring Vignettes 방법은 가상의 시나리오를 제시하여 응답자의 척도에 배치하도록 함으로써 응답자의 척도와 위치를 확인하고 원자료 배치값을 조정하는 방법이다.

특히, Aldrich-McKelvey(A-M) 방법은 최근에도 정치학에서 다양한 정치적 맥락을 연구하기 위해 A-M 스케일링이라는 이름으로 계속 사용하고 있다(예: Hare et al., 2014; Hollibaugh et al., 2013; Lo et al., 2014; Saiegh 2009). 또한 Anchoring Vignettes 방법은 World Health Survey(예컨대, 건강상태), Winsconsin Longitudinal Survey(예컨대, 이동성, 애착 등) 등의 설문조사 프로젝트에서 활용되었고(Grol-Prokopczyk et al., 2011), 이후 비모수적 방법과 모수적 방법으로 발전하였으며(King & Wand, 2007; Wand, 2013), 심리학 등 타 학문으로 확산되었다(Primi et al., 2016; He et al., 2017; Weiss & Roberts, 2018).

우선, 보건 및 건강 연구에서는 anchoring vignettes가 가장 활발하게 응용되는 분야 중 하나이다. Grol-Prokopczyk et al.(2011)은 Wisconsin Longitudinal Study 자료(n=2,625)를 활용하여 주관적 건강상태(general self-rated health)의 집단 간 차이를 평가했다. 앵커링 비네트를 활용하지 않은 모형에서는 여성이 남성보다 주관적 건강상태가 나은 것으로 나타났지만, 앵커링 비네트를 활용하여 조정한 모형에서는 이 차이가 사라졌다. 이를 통해, 여성은 남성에 비해 더 건강 낙관적(health-optimistic) 응답 스타일을 가지고 있음을 발견하였다.

정치학 분야에서 anchoring vignettes는 주로 정치적 태도, 이념, 효능감 등의 국가 간 비교를 위해 사용되어 왔다. King et al.(2004)은 앵커링 비네트 방법론을 처음 제안하면서 중국과 멕시코의 정치적 효능감(political efficacy)을 비교하였다. 기존의 방법으로는 중국인이 멕시코인보다 더 높은 정치적 효능감을 보였지만, 이는 직관에 반하는 결과였다. 이에 앵커링 비네트를 이용하여 응답 척도 이질성을 조정한 결과, 멕시코인이 중국인보다 더 높은 정치적 효능감을 가진 것으로 나타났다.

교육 연구에서는 anchoring vignettes는 학생의 자기보고 측정치의 비교 가능성을 향상시키기 위해 사용되어 왔다. He et al.(2017)은 PISA 2012에 참여한 64개 문화권의 15세 학생들을 대상으로 교사 지원(Teacher Support, TS)과 교실 관리(Classroom Management, CM)에 대한 학생 자기보고에 앵커링 비네트를 적용하였다. 연구결과에 의하면, 앵커링 비네트를 적용한 후, 문화권 간 학생 응답의 비교 가능성이 유의미하게 향상되었고, 교사 지원(CM)과 교실 관리 점수(TS)가 학업 성취도를 더 잘 예측하였다.

심리학 분야에서는 주로 성격 특성의 문화 간 비교를 위해 사용되어 왔다. Weiss &

Roberts(2018)은 르완다(n=423)와 필리핀(n=143)의 청소년 및 청년을 대상으로 성격의 5 요인(Big Five)에 대한 문화 간 비교에 앵커링 비네트를 적용하였다. 연구 결과, 앵커링 비네트 조정 점수가 자기보고 성격의 심리측정적 속성과 문화 간 비교 가능성을 향상시킨다는 점을 발견하였다.

이러한 선행 연구들의 흐름은 주관적 인식을 다루는 설문 조사에서 차별문항작용의 통제가 필요함을 시사한다. 이 연구는 이러한 선행 연구들의 논의를 경찰 신뢰라는 공공행정 영역으로 확장하여, 고정항목을 활용한 측정 방식이 사회과학 전반의 비교 연구에서 가질 수 있는 실천적 가치를 확인하고자 한다. 다음 장에서는 A-M 방법과 Anchoring Vinettes의 원리를 기반으로 '고정항목 방법(Anchoring Items)'을 제안하고자 한다. 이는 차별항목작용(DIF)이 존재하는 상황에서 개인 간의 신뢰할 수 있는 비교를 가능하게 하는 고정항목을 활용할 수 있는 방법이다. 경찰신뢰 수준을 비교하는 맥락에서, 경찰신뢰에 대한 원자료의 점수가 동일한 경우에도 고정항목를 활용하면 한 국가의 인지된 경찰신뢰 수준이 여전히 동일한지, 아니면 응답자들이 척도 범주를 해석하는 방식의 차이에서 실제 경찰신뢰 수준이 다른 것인지를 판단할 수 있게 된다. 다시 말해, 일본인과 영국인들의 경찰에 대한 신뢰 점수가 유사하다고 하더라도 응답자의 항목차별작용의 문제를 배제할 수 없다고 한다면, 국가 간 경찰신뢰의 비교에 있어서 항목차별작용을 고려한 경찰신뢰 점수의 조정을 대안적 측정방법으로 활용할 수 있다. 결국, '고정항목 방법'은 모든 개인을 경찰신뢰에 대한 공통된 인식과 가장 일치하는 위치로 신뢰척도를 조정함으로써, 차별문항작용으로 인한 문제를 해결하기 위한 노력이다.

4. 영국과 일본 간 경찰신뢰 비교에서 제기되는 차별문항작용의 문제

연구자가 영국과 일본의 경찰신뢰 수준을 비교하려 한다고 가정하자. 연구자는 영국과 일본이 국민을 대상으로 한 무작위표본추출을 실시하고, 경찰에 대해 신뢰하는지 동일한 문항의 설문지를 바탕으로 응답을 받게 될 것이다. 실제 세계가치조사(World Value Survey, 이하 WVS)는 1981년부터 최근까지 매년 세계 국가를 대상으로 국제 설문를 통해 경찰신뢰에 대한 질문을 하였다. WVS에서 활용한 경찰신뢰 관련 질문은 다음과 같다.

경찰에 대해 얼마나 신뢰하는지 말씀해 주시겠습니까? 매우 신뢰합니까, 상당히 신뢰합니까, 별로 신뢰하지 않습니까, 아니면 전혀 신뢰하지 않습니까?

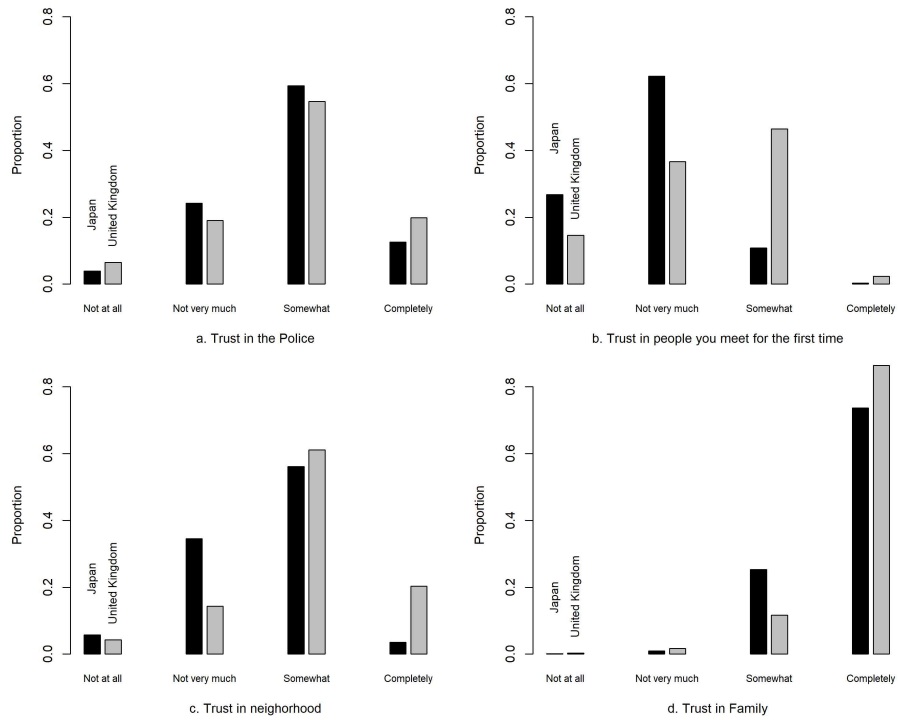
경찰신뢰 관련 자기 응답 질문과 함께, WVS 설문조사 응답자들은 이후에 동일한 순서 척도를 사용하여 고정항목으로 활용할 수 있는 다양한 집단에 대한 신뢰 관련 질문에 응답하도록 요청받았다. 고정항목의 신뢰 대상은 "처음 만나는 사람(people you meet for the first time)", "이웃 사람(neighborhood)", "가족(family)"이다. 이 고정항목 중에서 가장 덜 신뢰받는 대상은 "처음 만나는 사람"으로, 응답자들은 평균적으로 이 고정항목을 다른 고정항목 중 가장 최악의 신뢰대상으로 여긴다. 이 고정항목에 대한 일본과 영국 응답자들의 응답 분포는 〈그림 2〉과 같다.

〈그림 2〉는 실제 WVS에서 조사한 영국인과 일본인의 신뢰대상별 신뢰 응답의 분포를 보여주고 있다. 각 질문항에 대한 일본의 2010년과 2019년 응답비율, 그리고 영국의 2005년과 2022년 응답비율을 제시하고 있다. 영국과 일본은 모두 3번째 척도인 "상당히 신뢰한다"는 쪽에 속하는 응답자가 가장 많은 비율을 차지하고 있다. 이러한 결과는 〈표 1〉이 제시하는 함의를 그대로 내포하고 있다. 구체적으로 살펴보면, 〈그림 2〉에서 일본과 영국의 응답자들이 "신뢰 정도"에 대한 서수 척도를 다르게 사용하고 있다는 우려를 잘 보여주고 있다. 〈그림 2〉(a)에서 응답자들이 경찰신뢰에 대해 거의 유사한 분포를 보이고 있다. 하지만, 〈그림 2〉(b)에서 최악의 신뢰대상 "처음 만난 사람"에 대한 "신뢰정도"를 묻는 항목에 대한 평가에서 일본인보다 영국인들이 신뢰하는 정도에 확연한 차이를 보인다. 마찬가지로 〈그림 2〉(c)에서 중간 수준의 신뢰대상 "이웃"에 대한 "신뢰정도"를 묻는 항목에 대한 평가에서도 영국인들이 일본인 보다 신뢰하는 정도가 높다는 점을 재차 확인할 수 있다. 또한, 〈그림 2〉(d)에서 최상의 신뢰대상 "가족"에 대한 "신뢰정도"를 묻는 항목에 대한 평가에서는 영국인들은 일본인 보다 완전히 신뢰한다는 비율이 더욱 높다는 점을 확인할 수 있다. 따라서, 척도 사용 방식의 차이는 국가별 또는 문화권에 따라 체계적으로 다르게 나타날 수 있으며, 〈그림 2〉의 경우 일본은 고정항목 평가 시 척도의 낮은 쪽을 사용하는 경향이 더 강하다.

〈표 1〉에서 제시한 가상의 응답자 1은 영국인에 비교될 수 있고, 가상의 응답자 2는 일본인에 비교될 수 있다. 〈그림 2〉 패널 a에서 보는 바와 같이, 영국인과 일본인은 경찰신뢰에 대한 응답의 분포가 거의 유사하다. 하지만, 신뢰척도의 기준점에 대한 해석에 있어서 영국인들은 일본인에 비해서 보다 신뢰하는 경향이 높은 것으로 나타났다. 〈그림 2〉 패널 b에서 보는 바와 같이 처음 만난 사람에 대한 신뢰에 있어서 영국인들의 최빈응답은 "다소 신뢰한다"는 척도인데 비해, 일본인들의 최빈응답은 "신뢰하지 않는다"는 척도이다. 이러할 경우, WVS 원시자료에서 영국인과 일본인의 "경찰신뢰"는 유사하다고 하더라도, 그들의 신뢰성의 차이로 인해 신뢰 척도를 다르게 해석함으로 인한 오

염이 발생하였다고 가정한다면 영국인과 일본인의 "경찰신뢰"가 유사하다는 추론은 정당화될 수 없을 것이다.

<그림 2> 일본·영국의 경찰신뢰와 신뢰 관련 응답 분포



원 자료: WVS Time Series 1981~2022 Data

주: 영국 경찰신뢰(N=892, year = 2005, 평균=2.88), 일본 경찰신뢰(N=1,768, year = 2009, 평균 2.81)

Ⅲ. 고정항목을 활용한 비모수적 측정

1. 고정항목 활용의 전제 조건과 검증 가능한 가정

차별문항작용(DIF) 보정을 위해 고정항목을 활용하기 위해서는 다음의 네 가지 전제조건과 그에 수반되는 통계적 가정이 충족되어야 한다.

첫째, 측정 척도의 동일성과 항목 등가성(Item Equivalence) 가정이다. 고정항목과 자

기평가 항목(경찰 신뢰)은 응답 범주의 수가 동일해야 하며, 잠재적인 측정 스펙트럼이 일치해야 한다. 만약 자기평가 항목은 4점 척도인데 고정항목이 5점 척도라면, 동일한 심리적 기준점을 적용할 수 없기 때문이다. WVS 설문에서 본 연구가 선정한 고정항목 후보군(가족, 이웃, 타인 신뢰 등)과 분석 대상인 '경찰 신뢰' 항목은 모두 동일한 4점 리커트 척도로 측정되어 이 가정을 충족한다.

둘째, 응답 일치성(Response Consistency) 가정과 변동성 요건이다. 응답자가 자기평가 문항과 고정항목 문항에 대해 동일한 척도 사용 기준(Thresholds)을 적용한다는 가정이다. 이를 실증적으로 검토하기 위해 고정항목은 자기평가 항목보다 국가·집단 간 변동성이 낮아야 한다. 〈그림 3〉의 패널 a에서 확인되듯, 가족·이웃·타인 신뢰의 평균 밀도곡선은 경찰 신뢰에 비해 분포의 폭이 좁고 집중되어 있다. 이는 고정항목이 특정 국가의 특수성이나 정책 효과에 민감하게 반응하기보다, 척도 보정의 기준점(Anchors)으로서 안정적인 기능을 수행할 수 있음을 의미한다. 반면 패널 b의 정부기관 신뢰 항목들은 변동성이 커 고정항목으로 부적합하다.

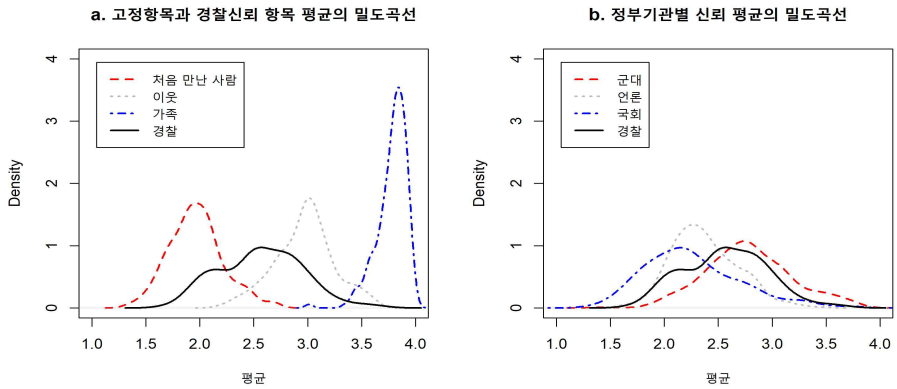
셋째, 측정 정보의 보존을 위한 최소 항목 수 요건이다. 비모수적 척도 변환 방식에서 변환 후의 정보 손실을 막기 위해 고정항목의 수(m)는 최소 3개 이상이어야 적절하다. C-Scale 변합 시 결과값은 $2m+1$개의 범주를 가지며, B-Scale은 $m+1$개의 구간을 형성한다. 고정항목이 3개인 경우($m=3$), C-Scale은 7점, B-Scale은 4점 척도로 변환되어 원자료(4점)의 측정 정보량을 유지하거나 확장할 수 있다. 만약 고정항목 수가 부족할 경우 변환 후 척도가 지나치게 단순화되어 분석의 정교함이 떨어지게 된다.

넷째, 문항 동등성(Vignette Equivalence) 가정과 서열 불변성 요건이다. 모든 응답자가 고정항목들 사이의 상대적 서열을 동일하게 인식한다는 가정이다. 즉, 국가나 문화에 상관없이 '가족 〉 이웃 〉 처음 만난 사람'이라는 신뢰의 위계가 엄격하게 유지되어야 한다. 〈그림 3〉 패널 a의 밀도 곡선이 서로 중첩되지 않고 명확한 순서로 분리되어 나타나는 것은 이러한 서열 불변성이 실증적으로 지지됨을 보여준다. 만약 특정 국가에서 이 순서가 역전된다면 고정항목으로서의 보편적 기준점 역할을 수행할 수 없으나, 이 연구의 대상 항목들은 전 국가적 차원에서 안정적인 위계를 유지하고 있다.

이러한 검증 결과를 종합해 볼 때, 본 연구에서 국가 간 경찰 신뢰를 측정 및 보정하기 위해 선정한 세 가지 고정항목(가족, 이웃, 처음 만난 사람)은 다음과 같은 측면에서 방법론적 적합성을 지닌다. 첫째, 해당 항목들은 응답자의 주관적 태도가 강하게 투영되는 자기평가 항목(경찰 신뢰)에 비해 국가 및 집단 간 변동성이 낮게 나타나, 척도 보정의 기준점으로서 필수적인 측정의 안정성을 확보하기에 유리하다. 둘째, 신뢰 대상에

따른 보편적 위계가 응답자들 사이에서 전 국가적으로 엄격하게 유지되고 있음이 확인되었다. 이는 문항 동등성 가정을 실증적으로 뒷받침하는 것으로,이 연구가 취하는 비모수적 보정 방식이 차별문항작용(DIF)에 따른 측정 편향을 교정하는 데 있어 타당한 도구로 기능할 수 있음을 시사한다.

〈그림 3〉 고정항목과 기관신뢰 항목 평균의 밀도곡선



원 자료: WVS Time Series 1981~2022 Data. 국가(108개)의 2~3개년 조사결과인 307개 표본의 평균임

2. 비모수적 "C" 척도 방법

고정항목 기반 비모수적 대인 비교 방법은 잠재적으로 비교 불가능한 자기 응답을 차별문항작용(DIF)이 제거된 새로운 측정값으로 대체하는 것이다. 이는 앵커링 비네트(Anchoring Vinettes)를 활용한 비모수적 비교방법과 유사하다. 고정항목 대상에 대한 개인의 평가는 자기 응답이 재조정되는 기준점 역할을 한다. 고정항목을 사용하는 일반적인 논리는 문항동등성과 응답일치성이라는 강한 가정을 요구하는 "C" 척도 방법이 있다. 이는 직관적으로 매력적이지만, 신뢰할 수 있는 비교를 생성하기 위해 불합리하고 지나치게 엄격한 가정을 전제한다는 비판이 제기될 수 있다(Wand, 2013).

"C" 척도는 자기 응답을 "C" 점수로 변화하는데, 다음의 정의에 따른다. C_i를 해당 고정항목 세트와 비교하여 측정하고자 하는 항목의 자기 응답으로 정의한다. y_i는 측정하고자 하는 항목에 대한 자기 평가 응답으로, $v_{i1}, \ldots, v_{im}$를 i번째 응답자의 m개 고정항목 응답이다. 모든 고정항목에 대해 일관된 순위를 매긴 응답자($v_m < v_{m+1}$)의 경

우에는, 다음과 같이 DIF 보정된 자기 응답 평가 C_i를 생성할 수 있다.

$$C_i = \begin{cases} 2m+1 & if \quad v_{im} < y_i < v_{i,m+1} \\ 2m & if \quad y_i = v_{im} \end{cases} \tag{3}$$

여기서 $m \in 1, ..., M$은 고정항목의 인덱스이고, $v_{i0} = -\infty$이며 $v_{i,\,M+1} = +\infty$이다. 따라서 측정하고자 하는 항목에 대한 개인의 속성을 측정하는 작업은 응답자가 자신을 하나 이상의 기준 대상보다 높거나, 낮거나, 또는 동일하게 평가하는지 여부를 확인하는 작업으로 재구성할 수 있다. 수식 (3)에서 고정항목 1과 고정항목 2에 대한 실제 평가가 동일하다면(예컨대, $y_i = v_{i,1} = v_{i,2}$), C_i는 스칼라 값이 아니라 벡터값을 갖게 되는데 $C_i \in \{2, 3, 4\}$가 된다. 이 경우 동점으로 인한 벡터를 어떻게 처리하느냐는 또 다른 문제이다.

수식 (3)을 〈표 1〉의 가상 응답자의 경우에 적용을 해보자. 즉, 가상 응답자 1의 C_1과 가상 응답자 2의 C_2를 구하는 것이다. 우선, 가상 응답자 1의 경찰신뢰에 대한 관측값(y_1)은 '2'이고, 가상 응답자 2의 경찰신뢰에 대한 관측값(y_2)도 '2'이다. 여기서 고정항목은 가족, 이웃, 처음 만난 사람 등 3개이므로, $m = 3$이다. 이제 가상 응답자 1은 경찰신뢰 관측값 2가 고정항목 1의 관측값과 같으므로($v_{11} = 2$), $C_1 = 2m = 2 \times 1 = 2$가 된다. 다음으로 가상 응답자 2는 경찰신뢰 관측값 2가 고정항목 2의 관측값과 같으므로($v_{22} = 2$), $C_2 = 2m = 2 \times 2 = 4$이다. 따라서, 〈표 1〉의 가상 응답자 1의 경찰신뢰 관측값과 가상 응답자 2의 경찰신뢰 관측값이 모두 2를 제시하고 있지만, 실제로는 고정항목과 비교할 때, 가상 응답자 1의 조정된 경찰신뢰는 2가 되는데 반해 가상 응답자 2의 조정된 경찰신뢰는 4가 된다.

그런데, King et al.(2004)에 의하면, "C"척도 방법은 기본적으로 고정항목 등가성과 응답일치성을 가정한다. 여기서, 고정항목 등가성(Vignette Equivalence)은 모든 응답자가 각각의 고정항목에 대한 실제 수준을 동일한 방식으로 이해하는 것을 의미한다. 즉, 모든 응답자가 잠재척도 상의 동일한 위치에서 각각의 고정항목을 인식해야 한다는 것을 의미한다. 여기서 동일한 위치라는 것은 고정항목의 순서가 동일하다는 것을 의미하는 것이고, 고정항목별 기준점(τ)의 위치가 동일해야 한다는 것은 아니다. 만약 차별문항작용이 존재하면, 고정항목 등가성이 충족되더라도 잠재척도가 달라지기 때문에 고정항목에 대한 관찰된 평가는 개인마다 다를 수 있다. 따라서, 〈표 1〉의 가상 응답자 1과

가상 응답자 2는 고정항목에 대한 잠재척도 상의 위치는 "처음 만난 사람", "이웃", "가족" 순으로 동일하지만, 고정항목별 기준점이 다르기 때문에 가상 응답자 1은 "처음 만난 사람"에 대해 "신뢰하지 않는다"고 평가하지만, 가상 응답자 2는 "매우 신뢰하지 않는다'고 평가하게 되는 것이다.

다음으로, 응답일치성(Response Consistency)은 한 개인에게 특정 설문이 주어진다면 그것이 고정항목이든 자기 평가 항목이든 잠재척도를 동일한 방식으로 활용하여 응답한다는 가정이다. 즉, 응답자가 고정항목과 자기 평가 질문을 평가할 때 동일한 기준점 위치를 사용해야 한다는 것을 의미한다. 응답일치성에 의하면, 차별문항작용은 응답자간에 발생하는 문제이지 개별 응답자에 대한 질문문항 간에 발생하는 문제는 아니다. 이러한 이유 때문에 응답일치성은 고정항목에 대한 개인의 평가를 자기 응답 평가와 연결한다는 점에서 고정항목 기반 분석방법의 기본 전제라고 할 수 있다.

하지만, 모든 사람이 각 기준 대상을 정확히 동일한 방식으로 인식할 것이라는 "고정항목 등가성" 가정은 지나치게 엄격하다. 응답자들이 동일한 고정항목 질문을 받더라도, 고정항목이 잠재 척도 상에서 어느 위치에 있는지에 대한 판단은 응답자마다 다를 수 있다. 즉, 〈표 1〉 가상 응답자들과는 달리, 어떤 응답자는 신뢰하는 대상의 순서를 "이웃", "처음 만난 사람", "가족" 순으로 위치시킬 수 있다. 따라서, 고정항목의 등가성 가정은 자기 평가에서 차별적 항목 기능(DIF)이 없다는 가정만큼이나 경험적으로 타당하지 않을 수 있다(Wand, 2013).

3. 비모수적 "B" 척도 방법

Wand(2013)는 King et al.(2004)이 제안한 비모수적 "C" 척도 보다 약한 가정 하에서도 신뢰할 수 있는 비교를 생성하는 비모수적 "B" 척도를 제안하였다. B 척도는 각 개인이 각 고정항목에서 인지하는 평균적인 위치를 기준으로 한 상대적인 위치로 정의한다(Wand, 2013).

$$B_i = m \;\; if \;\; \tilde{v}_{0,\,m-1} \leq \tilde{y}_i < \tilde{v}_{0m} \qquad (4)$$

여기서 $\tilde{v}_{0m}$은 실제 잠재척도의 위치 또는 해당 집단의 고정항목(m)에 대한 평균적인 잠재척도 위치($\tilde{v}_{0m} = E(\tilde{v}_{im})$)이다. 다시 말해, 기준점의 정의와 마찬가지로,

$\tilde{v}_{i0} = -\infty$ 이고, $\tilde{v}_{i,m+1} = +\infty$ 이다. 수식 (4)에서 고정항목 1과 고정항목 2에 대한 실제 평가가 동일하다면(예컨대, $y_i = v_{i,1} = v_{i,2}$), B_i는 스칼라 값이 아니라 벡터 값을 갖게 되는데 $B_i \in \{1, 2\}$가 된다. C 척도방법과 마찬가지로, 동점으로 인한 벡터를 어떻게 처리할 것인가에 대해 다양한 방법이 제시되고 있다. 간략히 살펴보면, 동점처리에 있어서 크게 (1) 고정항목 동점사례를 삭제하는 방법, (2) 주어진 백터구간 내에서 균등하게 배분하는 방법, (3) 절삭 순서화 프로빗 모형을 활용하여 배분하는 방법, (4) 최소 엔트로피 기준으로 배분하는 방법 등으로 나눌 수 있다.

수식 (4)을 〈표 1〉의 가상 응답자의 경우에 적용을 해보자. 즉, 가상 응답자 1의 B_1과 가상 응답자 2의 B_2를 구하는 것이다. 우선, 가상 응답자 1의 경찰신뢰에 대한 관측값(y_1)은 '2'인데, 이는 영국인의 첫 번째 고정항목의 기댓값(여기서는 가상의 영국인이 한 명이므로 그 사람의 값이 기댓값이 됨)은 $\tilde{v}_{01} = \tilde{v}_{11} = 2$이고, 두 번째 고정항목의 기댓값은 $\tilde{v}_{02} = \tilde{v}_{12} = 3$이며, 세 번째 고정항목의 기댓값은 $\tilde{v}_{03} = \tilde{v}_{13} = 4$이다. 그렇다면, 가상 응답자 1의 조정된 값은 $\tilde{y}_1 = 2$이고, 이는 첫 번째 항목값 1보다 크거나 같으므로 $B_1 = 2$이다. 일본인 가상 응답자 2의 첫 번째 고정항목의 기댓값은 $\tilde{v}_{02} = \tilde{v}_{12} = 1$이고, 두 번째 고정항목의 기댓값은 $\tilde{v}_{02} = \tilde{v}_{22} = 2$이며, 세 번째 고정항목의 기댓값은 $\tilde{v}_{03} = \tilde{v}_{23} = 3$이다. 그렇다면, 가상 응답자 2의 조정된 값은 $\tilde{y}_2 = 2$이고, 이는 두 번째 항목값 2보다 크거나 같으므로 $B_2 = 3$이 된다.

비록 $\tilde{y}_i$와 $\tilde{v}_{0m}$은 모두 응답자의 마음속에 숨어 있는 잠재 척도를 전제로 하기 때문에 관찰되지 않지만, 설문 응답을 바탕으로 이들의 상대적 위치를 파악할 수 있다. 여기서 B-척도의 장점은 기준점의 위치에 의존하지 않으므로, 결과적으로 고정항목 등가성이나 구간 등가성 없이도 상대적 비교가 가능하다는 점이다(Wand, 2013). 그럼에도 불구하고, 여전히 B 척도도 C 척도와 마찬가지로 응답일관성 가정을 전제한다. 모든 응답자들이 자기 평가를 포함하는 고정항목의 평가 순서가 잘못되어서는 안 된다는 것이다. 따라서, 〈표 1〉의 상황에서 응답자가 신뢰한다는 순서가 "가족", "처음 만난 사람", "이웃"과 같이 순서가 바뀌어서는 안된다.

4. 소결

이처럼 앵커링 비네트(Anchoring Vignettes)를 활용하여 DIF를 보정하려는 접근은 사회과학 전반에서 오랜 기간 논의되어 왔으나, 최근 연구들은 이 방법론이 전제하는 두

가지 핵심 가정인 '응답 일치성(Response Consistency)'과 '고정항목 등가성(Vignettes Equivalence)'의 충족 여부에 대해 보다 엄밀한 검증을 요구하고 있다(Wand, 2013). 특히 대규모 다수 국가 비교에서는 국가별로 고정항목(Anchoring Items)에 대한 이해도나 응답 범주 사용의 편향(극단 응답 성향, 중립 응답 성향 등)이 체계적으로 다르게 나타날 수 있다는 한계가 제기되기도 한다.

그럼에도 불구하고 이 연구에서 WVS 95개국이라는 방대한 데이터를 통해 고정항목 방법론을 제안하는 이유는, 앵커링 비네트 또는 문항반응이론(Item Response Theory)에 기반을 둔 복잡한 모수적 모델이 갖는 계산적 난해함을 극복하면서도 다양한 데이터에서 직관적으로 응답 기준의 이질성을 통제할 수 있는 비모수적 대안으로서의 가치가 있기 때문이다. 최근의 논의들은 완전한 보정보다는 보정 전후의 순위 변동성 자체를 하나의 국가별 문화적 지표로 해석하려는 시도로 확장되고 있으며, 이 연구 역시 이러한 맥락에서 전통적 측정치와의 상관관계 이면에 존재하는 '미세한 편향'을 포착하기 위한 시도이다.

Ⅳ. 고정항목을 활용한 국가간 경찰신뢰 비교

제3장에서는 국가 간 경찰신뢰를 비교할 때 각국 응답자의 주관적 신뢰 기준(DIF)을 통제하기 위한 대안으로 고정항목(Anchoring Items) 기반의 척도 변환 방식을 제안하였다. 이 장에서는 이 방법론적 대안을 실제 글로벌 데이터에 적용하여 그 타당성과 분석적 유용성을 실증적으로 검토하고자 한다.

1. 자료와 측정

이 연구에서 활용한 자료는 세계가치조사(World Values Survey)의 제7차 웨이브(2017-2022) 데이터이다. WVS는 전 세계 사회과학 연구자들에 의해 널리 활용되는 공신력 있는 국제 비교 설문 프로젝트로, 각 국가의 만 18세 이상 성인 남녀를 모집단으로 하여 확률표집방법(Probability Sampling)을 통해 표본을 추출하였다. 이 연구의 분석에 포함된 최종 국가는 총 95개국이고, 각 국가별 표본 크기는 대략 1,000명에서 3,000명 수준으로 구성되어 있다.

핵심 분석 변수인 경찰 신뢰 문항은 "당신은 다음에 나열된 조직들에 대해 얼마나 신뢰하십니까?"라는 공통 질문 아래, '경찰(The Police)' 항목에 대해 4점 척도(1: 매우 신

뢰함, 2: 어느 정도 신뢰함, 3: 별로 신뢰하지 않음, 4: 전혀 신뢰하지 않음)로 응답하도록 설계되었다. 이 연구에서는 분석의 편의와 직관적 해석을 위해 이를 역코딩하여 점수가 높을수록 경찰신뢰도가 높은 것으로 재구성하였다. 고정항목(Anchoring Items)으로 활용된 변수는 일반적 신뢰(General Trust) 항목들이다. 설문 대상자들은 '가족(Your family)', '이웃(Your neighborhood)', '처음 만난 사람(People you meet for the first time)'에 대해 각각 어느 정도 신뢰하는지를 동일한 4점 척도로 응답하였다. 이러한 세 가지 항목은 신뢰의 대상이 개인적 친밀도에 따라 계층화되어 있어, 응답자의 주관적 척도 기준점을 파악하는 비모수적 보정 도구로서 적절한 요건을 갖추고 있다.

자료의 결측치 처리와 관련하여서는, 경찰 신뢰 및 세 가지 고정항목 중 하나라도 무응답인 사례는 분석 대상에서 제외하였다. 각 국가별 조사연도, 표본 수, 원점수, 변환점수 통계량 등을 부록에 상세히 수록하였다(〈부록1〉 참조). 특히, 본 연구는 국가별 순위 비교의 안정성을 검증하기 위해 비모수적 부트스트랩(Non-parametric Bootstrap) 기법을 적용하였다. 이를 위해 각 국가별로 200회의 복원 재표집(Resampling)을 실시하였으며, 매 회차마다 척도 변환 과정을 독립적으로 재실행하여 95% 신뢰구간을 산출하였다.

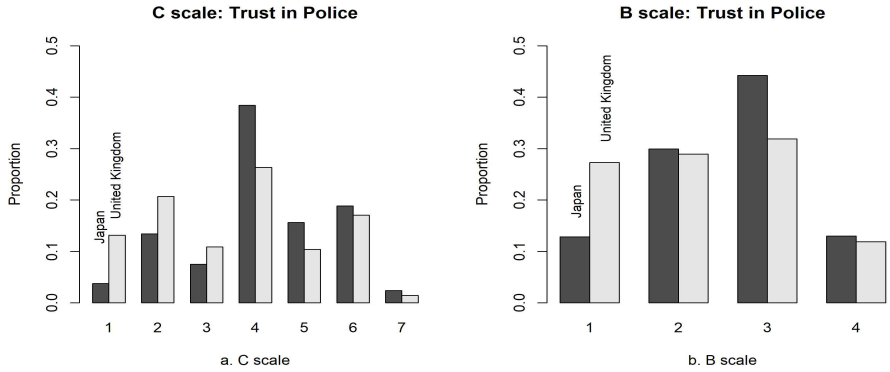
2. 영국과 일본 간의 경찰신뢰 비교

국가 간 경찰신뢰 비교 연구에서 문화적 차이에 따른 차별문항작용(DIF)이 존재한다면, 고정항목을 활용한 척도 변환이 원점수보다 실제 현상을 더 정확히 반영할 가능성이 크다. 이를 확인하기 위해 영국(2005년)과 일본(2009년)의 사례에 고정항목 보정법(B 및 C 척도)을 적용하여 그 분포의 변화를 분석하였다(〈그림 2〉과 〈그림 4〉 참조).

원자료 상에서 영국(2.88점)과 일본(2.81점)의 경찰신뢰 평균은 영국이 미세하게 높은 유사한 수준으로 나타났다. 그러나 고정항목을 활용하여 재척도화한 결과, 놀라운 순위 역전 현상이 관찰되었다. C 척도 기준 영국은 3.57점으로 하락한 반면 일본은 4.15점으로 상승하며 신뢰 수준이 역전되었고, B 척도 역시 동일한 경향성을 보였다.

이는 영국인들이 일본인에 비해 전반적인 '신뢰 성향' 자체가 높은 경향이 있음을 시사한다. 영국 응답자들에게 경찰은 다른 신뢰 대상(이웃 등)에 비해 상대적으로 낮은 위치에 있으나, 일본 응답자들은 경찰을 이웃과 비슷하거나 더 높은 수준으로 신뢰하고 있었다. 즉, 원점수 기반 비교가 간과했던 '국가별 응답 기준점의 차이'가 고정항목 보정을 통해 드러나면서 보다 타당한 비교가 가능해진 것이다.

〈그림 4〉 일본과 영국의 경찰신뢰에 대한 비모수적 측정



원 자료: WVS Time Series 1981~2022 Data

주 1: 영국 경찰신뢰(N=892, year = 2005, 원자료 평균 = 2.88, C scale 평균 = 3.57, B scale 평균 = 2.28), 일본 경찰신뢰(N=1,768, year = 2009, 원자료 평균 = 2.81, C scale 평균 = 4.15, B scale 평균 = 2.57)

주 2: 고정항목 B, C Scale 변환과정에서 고정항목의 동점(Ties) 처리는 "균등배당(Uinform Allocation)" 방식 적용

3. WVS 국제프로젝트의 경찰신뢰에 대한 국가간 비교

WVS 시계열 데이터(1981-2022)는 전 세계 108개국을 대상으로 실시된 총 307개의 국가-연도(Country-Year) 관측치로 구성되어 있다. 본 연구에서는 국가 간 비교의 최신성을 확보하기 위해 각 국가의 가장 최근 조사 회차(Most Recent Wave) 자료를 추출하였으며, 분석의 핵심인 3종의 고정항목(가족, 이웃, 처음 만난 사람) 조사가 누락된 13개국을 제외한 최종 95개국을 분석 대상으로 확정하였다. 추출된 데이터에 대해 경찰 신뢰 원점수 평균과 고정항목 보정을 거친 C 척도 및 B 척도 변환 평균 점수를 산출하였으며, 이를 바탕으로 국가별 상대적 신뢰 수준과 순위 변화를 분석하였다(부록 〈표 1〉 참조). 이러한 분석 결과를 해석하면 다음과 같다.

첫째, 원점수와 변환된 점수(C척도 또는 B척도) 간에 높은 상관관계를 확인할 수 있으나, 국가 간 순위를 크게 역전시킬 수 있을 정도로 변환된 점수는 원점수와 차이를 보이는 경우도 상당하다. 예컨대, 핀란드는 2005년 조사에서 원점수는 3.26으로 5위를 차지하였으나 B척도에서는 2.49, C척도에서는 3.99점으로 23위로 나타나 순위 측면에서 크게 하강하였다. 한편, 니카라과는 2020년 조사에서 원점수는 2.05으로 87위에 해당하였으나 B척도에서는 2.24, C척도에서는 3.47점으로 46위로 나타나 순위가 크게 상승하였다.

둘째, 변환된 C척도, 변환된 B척도의 국가별 평균점수는 상관계수가 0.99에 이를 정도로 거의 유사한 결과를 나타내고 있다(부록 〈그림 1〉 패널b). 이는 C척도와 B척도 변환이 개인적 수준에서 가정을 달리하고 다르게 점수를 변환한다고 하더라도, 국가단위로 집계된 데이터에서 C척도 변환과 B척도의 변환은 큰 차이를 초래하지 않는다는 것이다.

특히, 같은 수준에 있던 국가들이 고정항목을 활용한 점수변환 결과가 원점수의 수준과 크게 달라지는 두 가지 사례를 살펴보면 다음과 같다. 먼저, 경찰신뢰의 원점수가 고정항목을 적용하였을 때 크게 하락한 경우로 네덜란드의 사례를 원점수가 비슷하였던 그리스와 비교하였다(부록 〈그림 2〉 참조). 패널 a에서 그리스와 네덜란드의 원점수의 분포를 보여주고 있다. 그리스와 네덜란드는 원점수의 분포에 있어서 거의 유사하고, 원점수 평균이 각각 2.82와 2.79점으로 매우 유사하다. 하지만, 고정항목에 대한 분포(패널 b, 패널 c, 패널 d)를 살펴보면, 그리스와 네덜란드는 응답자의 신뢰성향 측면에서 상당한 차이를 보여주고 있다. 네덜란드 응답자들은 그리스 응답자에 비해 "처음 만난 사람", "이웃"에 대한 신뢰가 더욱 높은 것으로 나타났다. 네덜란드 응답자들은 고정항목이 높게 형성되어 있기 때문에, 경찰신뢰의 원점수가 그리스 응답자와 거의 유사하다고 하더라도 고정항목을 활용한 척도 변환을 적용할 경우, B점수와 C점수 모두가 그리스 응답자와 비교하여 낮게 형성되는 것이다(패널 e, 패널 f 참조). 결국, 네덜란드의 경찰신뢰 원점수 순위가 35위 이었던 것을, 고정항목을 활용한 척도 변환으로 인해 85위 내지 86위로 하락하게 만든 것이다.

다음으로, 경찰신뢰의 원점수가 고정항목을 적용하였을 때 크게 상승한 경우로 에콰도르의 사례를 원점수가 비슷하였던 불가리아와 비교하였다(부록 〈그림 3〉 참조). 패널 a에서 불가리아와 에콰도르의 원점수의 분포를 보여주고 있다. 불가리아와 에콰도르는 원점수의 분포에 있어서 거의 유사하고, 원점수 평균이 각각 2.53와 2.55점으로 매우 유사하다. 하지만, 고정항목에 대한 분포(패널 b, 패널 c, 패널 d)를 살펴보면, 불가리아와 에콰도르는 응답자의 신뢰 성향 측면에서 상당한 차이를 보여주고 있다. 에콰도르 응답자들은 불가리아 응답자에 비해 "처음 만난 사람", "이웃"에 대한 신뢰가 매우 낮은 것으로 나타났다. 즉, 에콰도르 응답자들은 고정항목이 상대적으로 낮게 형성되어 있기 때문에, 경찰신뢰의 원점수가 불가리아 응답자와 거의 유사하다고 하더라도 고정항목을 활용한 척도 변환을 적용할 경우, B점수와 C점수 모두가 불가리아 응답자와 비교하여 높게 형성되는 것이다(패널 e, 패널 f 참조). 결국, 에콰도르의 경찰신뢰 원점수 순위가 53위 이었던 것을, 고정항목을 활용한 척도 변환으로 인해 12위로 상승하게 만든 것이

다.

4. 보정 전후 변동폭이 큰 국가와 적은 국가의 특성 비교

앞서 살펴본 바와 같이 95개국 전체 데이터를 대상으로 한 상관관계 분석은 두 측정 방식 간의 높은 일관성을 보여주지만, 개별 국가 차원에서 발생하는 미세한 순위 변동은 국가 간 비교의 타당성을 재고하게 만드는 중요한 단초를 제공한다. 특히 특정 국가군에서 나타나는 유의미한 순위 변화는 차별문항작용(DIF)이 단순히 통계적 오차에 그치는 것이 아니라, 각국의 문화적 응답 성향과 깊게 연관되어 있음을 시사한다. 이러한 맥락에서 보정 전후의 변동폭이 두드러지게 나타난 국가와 그렇지 않은 국가를 대비하여 그 기저에 깔린 응답 특성을 구체적으로 살펴보고자 한다.

분석 결과(〈표 2〉, 부록 〈그림 4〉 참조)에 따르면, 고정항목 보정 전후의 변동폭은 국가별로 뚜렷한 차이를 보였다. 네덜란드(2022), 노르웨이(2007), 스웨덴(2011), 핀란드(2005)와 같은 북유럽 국가들은 전통적 원점수 방식과 달리 고정항목 보정 후 크게 변동하는 특성을 보였다. 이는 해당 국가 응답자들이 설문 척도를 사용할 때 매우 엄격한 주관적 기준점을 적용하고 있음을 시사한다(부록 〈그림 4〉 참조). 즉, 이들은 실제 신뢰 수준이 높음에도 불구하고 척도 상에서는 보수적인 점수를 부여하고 있었으며, 고정항목을 통해 이러한 응답 편향을 제거했을 때 비로소 실제의 높은 신뢰도가 순위로 표출된 것이다.

〈표 2〉 보정 전후 변동폭이 큰 국가와 적은 국가의 특성 비교

연번	변동폭 적은 국가	n	원 점수		B scale		C Scale	
			평균	순위	평균	순위	평균	순위
1	Moldova (2006)	1,008	1.97	88	1.96	84	2.91	84
2	Guatemala (2020)	1,011	1.84	91	1.85	92	2.69	91
3	Azerbaijan (2011)	961	2.64	44	2.59	13	4.19	13
4	Colombia (2018)	1,440	2.16	83	2.22	49	3.43	49
5	Brazil (2018)	1,557	2.46	66	2.40	31	3.80	32
연번	변동폭 큰 국가	n	원 점수		B scale		C Scale	
			평균	순위	평균	순위	평균	순위
1	Netherlands (2022)	1,689	2.79	35	1.93	86	2.87	85
2	Norway (2007)	999	3.05	15	2.13	65	3.25	65
3	Sweden (2011)	1,082	2.93	20	2.18	54	3.35	52
4	Rwanda (2012)	1,321	2.76	36	2.09	70	3.17	70
5	Finland (2005)	972	3.26	5	2.49	23	3.99	23

원 자료: WVS Time Series 1981~2022 Data
주 1: 변동폭 산식은 "(원점수 - 변환 scale)/원점수"임
주 2: 고정항목의 동점(Ties) 처리는 "균등배당(Uinform Allocation)" 방식 적용

반면 몰도바(2006), 과테말라(2020), 콜롬비아(2018), 브라질(2018) 등은 보정 전후의 변동이 거의 나타나지 않았다. 이들 국가는 고정항목에 대한 신뢰 기준과 경찰 신뢰에 대한 응답 기준이 일치하거나, 응답자들이 척도 전체를 활용하기보다 관대한 범주에 집중하여 응답하는 경향을 보였다. 이러한 결과는 DIF 보정 방법론이 모든 국가에서 동일한 효과를 발휘하기보다는, 북유럽 국가들과 같이 응답 기준이 유독 엄격한 문화권을 비교 대상에 포함할 때 측정의 타당성을 높여줄 수 있음을 실증적으로 보여준다.

종합하건대, 전통적 원점수와 보정점수(B척도 또는 C척도) 간에 높은 상관관계가 관찰됨에도 불구하고, 특정 국가군에서는 순위의 역전이나 수치의 유의미한 조정이 발생하였다. 이러한 차이가 발생하는 '임계 지점(Threshold)'을 구체적으로 살펴보면, 주로 고정항목(가족, 이웃, 타인)에 대한 응답이 특정 범주에 과도하게 쏠려 있거나, 응답자들의 고정항목 간 서열 역전 빈도가 높은 국가들에서 DIF 보정의 효과가 높게 나타났다.

예를 들어, 경찰 신뢰 원점수는 유사하나 고정항목에 대한 신뢰 기준이 매우 엄격한 국가(예: 그리스)와 상대적으로 관대한 국가(예: 동남아시아 일부 국가)를 비교할 때, 고정항목을 활용한 척도 변환은 단순 평균이 간과했던 '체감 신뢰의 밀도'를 재조정하는 역할을 한다. 따라서 이 연구에서 제안하는 방법론의 유용성은 모든 국가에서 동일하게 나타나기보다는, 응답 성향의 극단성이 강하거나 문화적 척도 기준의 편차가 큰 국가들을 비교 대상에 포함할 때 실질적인 분석적 통찰을 제공한다고 볼 수 있다. 이는 DIF 보정이 단순히 수치를 바꾸는 도구가 아니라, 비교 가능성의 한계를 식별하고 데이터의 신뢰 구간을 설정하는 엄밀한 진단 도구로서 기능할 수 있음을 시사한다.

5. 고정항목을 활용한 측정방법의 타당성과 강건성 검증

이 연구에서 제안한 고정항목을 활용한 척도 변환 방식은 모든 응답자가 고정항목의 서열을 동일하게 인식한다는 문항 동등성 가정을 전제로 한다. 그러나 문화권에 따라 가족, 이웃, 처음 만난 사람에 대한 신뢰의 의미가 다르게 해석될 가능성을 완전히 배제하기 어렵다. 이에 이 연구에서 제안한 방법론의 타당성을 검토하기 위해 외부 준거변수를 활용한 측정 타당성 검증과 민감도 분석을 수행하였다(〈표 3〉 및 부록 〈표 2〉 참조).

먼저, 고정항목 활용의 측정 타당성 검증이다. 변환 점수가 원점수보다 경찰 신뢰라는 구성 개념을 더 정확히 반영하는지 확인하기 위해, 외부 객관 지표인 부패인식지수

(CPI)와의 관련성을 개인 및 국가 수준에서 비교 분석하였다(부록 〈표 2〉 참조). 부패인식지수(CPI: Corruption Perceptions Index)는 국제투명성기구(Transparency International)에서 매년 발표하는 지표로, 공공부문의 부패에 대한 전문가 및 기업인의 인식을 0점(매우 부패)에서 100점(매우 청렴) 사이의 점수로 산출한 것이다. 분석에서는 WVS 조사 시점과 가장 인접한 연도의 국가별 CPI 데이터를 매칭하여 분석에 활용하였다. 분석 결과, 개인 수준의 경찰 신뢰 원점수와 CPI 간에는 강한 정(+)의 상관관계(ρ = 0.467, p-value 〈 .001)가 존재함을 확인하여 측정 도구의 기초적 준거 타당성을 확보하였다. 그러나 국가별 고정항목을 활용한 척도와 CPI와의 상관관계는 ρ = 0.25(p-value 〈 .05)로 감소하는 양상을 보였다. 일반적으로 상관계수의 수치가 낮아지는 것을 측정 효율성의 저하로 오인할 수 있으나, 국가 간 비교 연구의 특성을 고려할 때 이는 오히려 보정의 타당성을 입증하는 결과로 해석된다. 원점수 기반의 상관관계에는 각 국가의 문화적 응답 성향(예: 관대함이나 엄격함)이 두 변수 모두에 공통적으로 혼입되어 나타나는 '허위 상관(Spurious Correlation)'이 포함되어 있을 가능성이 크다. 따라서 고정항목 보정 과정을 거치며 상관계수가 낮아지는 현상은, 측정 도구 내에 잠재되어 있던 문화적 노이즈가 제거되고 경찰 신뢰와 부패 인식 사이의 '순수한 상관성'이 도출되는 과정으로 이해되어야 할 것이다.

다음으로 고정항목 활용 척도의 민감도 분석을 세 가지 방법으로 실시하였다(〈표 3〉 참조). 첫째, 고정항목 중 하나를 순차적으로 제외하여 B척도를 재산출하는 Leave-one-out 분석을 실시하였다. 고정항목의 구성 변화와 동점처리 방식의 변동에도 불구하고 보정 점수는 원점수와 매우 유의미한 상관관계(ρ = 0.77~0.81, p-value 〈 0.001)를 유지하는 것으로 나타났다. 특히 세 개의 고정항목 중 특정 항목(예컨대, '처음 만난 사람', '이웃', '가족')을 제외하더라도 상관계수의 변동폭이 미미하게 나타난 점은, 본 연구에서 선정한 고정항목들이 특정 문항의 편향에 휘둘리지 않고 안정적인 보정 기능을 수행하고 있음을 뒷받침한다.

둘째, 동점 처리(Tie-breaking) 방식에 따른 민감도를 점검하였다. 이 연구는 기본적으로 응답자가 고정항목과 자기평가 항목에 동일한 점수를 부여할 경우 그 범주 내에서 균등 배정하는 방식을 택하고 있다. 이를 '고정항목 동점 사례 삭제' 방법으로 변경하여 재분석한 결과, 원점수와의 상관계수는 0.62(p-value 〈 0.001)로 나타났다. 이는 서열 판단이 모호한 사례들을 분석에서 제외하더라도, 보정된 척도가 원본 데이터가 가진 국가 간 신뢰의 상대적 순위를 유의미하게 보존하고 있음을 의미한다.

셋째, 국가 간 순위 변동폭을 분석한 결과, 고정항목을 하나씩 제외하였을 때의 평균

순위 변동폭은 3.3위에서 4.7위 사이로 나타났다. 이는 95개국 전체 순위 범위를 고려할 때 매우 미미한 수준으로, 이 연구의 고정항목 구성이 특정 문항에 과도하게 의존하지 않는 강건성을 갖추었음을 의미한다. 반면, 동점 사례를 삭제하는 엄격한 조건을 적용했을 때는 변동폭이 7.2위로 다소 확대되었는데, 이는 설문 응답의 동점 구간이 국가 간 신뢰 수준의 미세한 차이를 식별하는 데 중요한 정보를 포함하고 있음을 시사한다. 따라서 이 연구가 채택한 균등 배분 방식이 정보 손실을 최소화하면서도 측정의 정교함을 높이는 타당한 선택이었음을 알 수 있다

〈표 3〉 고정항목 구성 및 분석 방식 변화에 따른 점수 상관관계 분석

분석 조건	원안 (3개 항목)	항목 1 (외인) 제외	항목 2 (이웃) 제외	항목 3 (가족) 제외	동점 처리 변경
원점수와의 상관성 (spearman's ρ)	0.80 ***	0.78 ***	0.81***	0.77 ***	0.62 ***
국가 간 순위 변동폭(평균)	-	4.7	3.5	3.3	7.2

원 자료: WVS Time Series 1981~2022 Data

주 1) p-value * 〈 0.05, ** 〈 0.01, *** 0.001

주 2) 분석 대상 표본 수(N): 95개 국가에 대한 분석으로 원안 및 민감도 분석 조건 1~3은 결측치를 제외한 전체 응답자 수(212,129명)를 유지하였으나, 동점 처리 변경의 경우 서열 판단이 불가능한 동점 사례가 제외되어 유효 표본 수가 약 60,133명으로 감소함

주 3) 국가 간 순위 변동폭은 분석조건별 국가들의 순위(Rank)를 매긴 후, 그 순위의 차이(절대값)를 계산하여 산술평균함

주 4) 동점처리는 (1) 주어진 백터구간 내에서 균등하게 배분하는 방법에서 (2) 고정항목 동점사례를 삭제하는 방법으로 변경한 결과임

V. 결론 및 시사점

이 연구는 King et al.(2004)이 제기한 차별문항작용(Differential Item Functioning, 이하 DIF)의 문제 인식을 바탕으로, 국가 간 비교 연구에서 발생하는 주관적 응답 기준의 이질성을 극복하기 위한 대안으로 '고정항목(Anchoring Items)'을 활용한 측정 방식을 제안한다. 국가 간 비교 연구에서 각국의 문화적·제도적 특성에 따라 응답자가 리커트 척도(Likert scale)에 부여하는 심리적 기준점이 다르다는 점을 인정한다면, 전통적 측정법으로 산출된 동일한 수치라도 실질적인 의미는 상이할 수 있음을 이해해야 한다.

예컨대, 경찰 신뢰도를 측정할 때 문화적 특성에 따라 모든 대상에 관대한 점수를 부여하는 응답 성향이 있는 국가와, 대단히 엄격한 잣대로 낮은 점수를 부여하는 경향이 있는 국가가 존재할 수 있다. 이러한 응답 기준의 이질성(Response Heterogeneity)을 통

제하지 못한 채 계산된 단순 평균값이 두 국가 간에 동일하게 나타났다고 해서, 이를 실질적인 신뢰 수준이 같다고 해석하는 것은 타당성 측면에서 심각한 오류를 범할 수 있다. 이는 설문 문항의 척도가 응답자의 배경에 따라 서로 다르게 기능하는 DIF의 본질적인 문제를 여실히 보여주는 사례이다.

이러한 측정상의 한계를 극복하기 위해 King et al.(2004)이 제안한 앵커링 비네트(Anchoring Vignettes)의 보정 원리를 응용하되, 설문 내 기존 항목을 '고정항목(Anchoring items)'으로 활용하는 변형된 접근법을 취한다. 이는 가상의 시나리오(Vignettes)를 직접 제시하는 전통적 방식에서 나아가, 응답자군 간 해석의 편차가 적고 보편적 위계성을 지닌 기존 설문 문항들을 척도의 기준점(Anchors)으로 재해석하는 방식이다. 구체적인 보정 절차로는 문항 동등성(Vignette Equivalence)과 응답 일치성(Response Consistency)의 엄격한 가정을 전제로 상대적 서열을 산출하는 비모수적 C-척도 변환 방법과 이보다 완화된 가정을 바탕으로 응답 범주의 불확실성을 구간으로 처리하는 B-척도 변환 방법을 제안하였다. 이 과정에서 모든 응답자가 고정항목(가족-이웃-타인)의 신뢰 위계에 동의한다는 '서열적 안정성'을 핵심 가정으로 설정하였다.

나아가 WVS(World Values Survey) 데이터를 바탕으로 전 세계 95개국에 대해 전통적 측정법과 고정항목 보정법을 각각 적용하여 비교 분석하였다. 분석 결과, 원자료 기반의 수치와 보정된 수치 간에는 전반적인 상관관계가 존재하나, 특정 국가들에서는 두 측정치 간에 현격한 순위 변동이 나타났다. 이러한 차이는 해당 국가의 응답 과정에서 DIF가 강하게 작용했음을 시사하며, 단순 평균값에 의존한 기존 해석에 기존 해석에 신중한 접근이 필요함을 시사한다. 또한, 연구가 제시한 방법론은 DIF를 해소하기 위한 실천적 대안으로서 의의가 있을 뿐만 아니라, 국가별 경찰 신뢰의 위치를 고정항목이라는 기준점과 비교하여 해석함으로써 보다 실질적이고 시사점 있는 분석을 가능케 한다. 이는 국가 간 경찰 신뢰의 진실에 한 걸음 더 다가가는 도구가 될 것으로 기대한다. 다만, 국가 간 경찰 신뢰의 변동성은 경찰 활동의 효과성, 경제적 상황, 부패 지수 등 객관적 요소에 의해 상당 부분 결정되므로, DIF에 기인한 변동성은 전체 변동성의 일부분임을 인지해야 한다. 또한 고정항목 보정법은 집단 간 '상대적' 위치를 정교화하는 데 유용하나, 이를 '절대적' 수치로 오인해서는 안 된다. 따라서 이 연구에서 제안하는 방법론은 전통적 측정법을 대체하기보다 이를 보완하고 교정하는 강건성 확인의 도구로 활용되는 것이 타당하다.

이 연구에서 제안한 고정항목을 활용한 비모수적 측정 방법은 국가 간 비교뿐만 아니라 단일 국가 내의 다양한 집단 간 비교나 시계열 분석에서도 활용이 가능하다. 차별문

항작용(DIF)은 단순히 국경을 넘어서는 문화적 차이에서만 발생하는 것이 아니라, 동일 국가 내에서도 응답자의 인구통계학적 특성이나 시대적 배경에 따라 응답 척도를 해석하는 기준이 달라질 때 발생하기 때문이다. 첫째, 단일 국가 내 집단 간 비교에 적용할 수 있다. 예를 들어, 세대 간(청년층과 노년층), 교육 수준 간, 혹은 성별 간에 특정 개념(예: 직무만족, 조직몰입, 삶의 만족도, 공정성 등)을 바라보는 주관적 기준점이 다를 경우, 고정항목을 통해 각 집단의 응답 기준을 정규화함으로써 보다 객관적인 비교가 가능하다. 둘째, 시계열 비교에서도 활용이 가능하다. 동일한 설문 문항이라 하더라도 사회적 분위기나 특정 사건의 발생 전후로 응답자가 체감하는 척도의 무게감이 달라질 수 있다. 이때 고정항목이 포함된 패널 설문조사를 지속적으로 실시한다면, 시간의 흐름에 따른 단순한 수치 변화가 아닌 실질적인 태도 변화를 추적하는 데 기여할 수 있다. 다만, 이러한 확장을 위한 핵심적인 적용 조건은 사용되는 고정항목이 비교 대상이 되는 모든 집단이나 시점에서 '서열적 안정성'을 유지해야 한다는 점이다. 즉, 이 연구에서 활용된 가족, 이웃, 처음 만난 사람에 대한 신뢰 수준의 위계가 집단이나 시간에 관계없이 일관되게 나타나야 한다. 만약 특정 집단에서만 고정항목의 서열이 뒤바뀐다면 이는 고정항목으로서의 기능을 상실하게 된다. 결론적으로 이 방법론은 국가 간 비교라는 맥락에서 그 효과가 극대화되지만, 응답 기준의 이질성(Response Heterogeneity)이 예상되는 모든 사회과학적 비교 연구로 확장될 수 있을 것이다.

마지막으로, 고정항목을 활용한 방법론은 설문지에 3개 이상의 적절한 고정항목이 존재해야 한다는 제약이 있다. 향후 연구에서는 이러한 한계를 넘어 각 국가 및 집단별 기준점을 모수적으로 추정하여 활용하는 방식을 통해 방법론적 정교화를 시도할 필요가 있으며, 이는 향후 과제로 남겨둔다.

《참 고 문 헌》

장현석. (2014). 경찰 신뢰도에 대한 한국과 일본 비교연구. 한국경찰연구, 13(2), 311-340.

정보성 · 이창배. (2018). 경찰신뢰의 영향 요인에 대한 다수준적 접근: 도구적 시각과 표현적 시각을 중심으로. 치안정책연구, 32(3), 7-44.

Aldrich, J. H., & McKelvey, R. D. (1977). A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections. American Political Science Review, 71(1), 111-130.

Bittner, E. (1970). The functions of the police in modern society: A review of background factors, current practices, and possible role models.

Cao, L., Lai, Y. L., & Zhao, R. (2012). Shades of blue: Confidence in the police in the world. Journal of criminal justice, 40(1), 40-49.

Cao, L., & Zhao, J. S. (2005). Confidence in the police in Latin America. Journal of criminal justice, 33(5), 403-412.

Grol-Prokopczyk, H., Freese, J., & Hauser, R. M. (2011). Using anchoring vignettes to assess group differences in general self-rated health. Journal of health and social behavior, 52(2), 246-261.

Hamm, J. A., Trinkner, R., & Carr, J. D. (2017). Fair process, trust, and cooperation: Moving toward an integrated framework of police legitimacy. Criminal justice and behavior, 44(9), 1183-1212.

Hare, C., Armstrong, D. A., Bakker, R., Carroll, R., & Poole, K. T. (2015). Using Bayesian Aldrich McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions. American Journal of Political Science, 59(3), 759-774.

He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. Journal of Cross-Cultural Psychology, 48(3), 319-334.

Hollibaugh, G. E., Rothenberg, L. S., & Rulison, K. K. (2013). Does it really hurt to be out of step?. Political Research Quarterly, 66(4), 856-867.

Jackson, J., Bradford, B., Stanko, B., & Hohl, K. (2012). Just authority?: Trust in the police in England and Wales. Willan.

Jang, H., Joo, H. J., & Zhao, J. S. (2010). Determinants of public confidence in police:

An international perspective. Journal of criminal justice, 38(1), 57-68.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. American political science review, 98(1), 191-207.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. Political Analysis, 15(1), 46-66.

Lo, J., Proksch, S. O., & Gschwend, T. (2014). A common left-right scale for voters and parties in Europe. Political Analysis, 22(2), 205-223.

Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016). Anchoring vignettes. European Journal of Psychological Assessment.

Saiegh, S. M. (2009). Recovering a basic space from elite surveys: Evidence from Latin America. Legislative Studies Quarterly, 34(1), 117-145.

Schaap, D., & Scheepers, P. (2014). Comparing citizens' trust in the police across European countries: An assessment of cross-country measurement equivalence. International Criminal Justice Review, 24(1), 82-98.

Sunshine, J., & Tyler, T. R. (2003). The role of procedural justice and legitimacy in shaping public support for policing. Law & society review, 37(3), 513-547.

Transparency International. (2025). Corruption Perceptions Index. Retrieved from https://www.transparency.org/en/cpi/

Wand, J. (2013). Credible comparisons using interpersonally incomparable data: Nonparametric scales with anchoring vignettes. American Journal of Political Science, 57(1), 249-262.

Weiss, S., & Roberts, R. D. (2018). Using anchoring vignettes to adjust self-reported personality: A comparison between countries. Frontiers in psychology, 9, 325.

[부록]

〈표 1〉 WVS 95개 국가의 경찰신뢰 비교

연번	국가(조사연도)	표본수	원 점수				B scale				C Scale			
			평균	하한	상한	순위	평균	하한	상한	순위	평균	하한	상한	순위
1	Algeria(2014)	918	2.63	2.52	2.66	46	2.42	3.61	3.86	29	3.85	2.31	2.43	29
2	Andorra(2018)	981	2.88	2.83	2.93	24	2.46	3.82	4.00	27	3.92	2.41	2.50	27
3	Argentina(2017)	918	2.08	2.03	2.15	86	1.75	2.42	2.66	94	2.49	1.72	1.84	94
4	Armenia(2021)	1,140	2.32	2.25	2.38	78	2.05	2.98	3.19	75	3.09	1.99	2.09	75
5	Australia(2018)	1,680	3.10	3.07	3.13	12	2.53	3.98	4.13	19	4.06	2.49	2.56	19
6	Azerbaijan(2011)	961	2.64	2.59	2.71	44	2.59	4.07	4.29	13	4.19	2.53	2.65	13
7	Bangladesh(2018)	969	2.43	2.39	2.51	71	2.01	2.94	3.13	79	3.02	1.97	2.06	79
8	Belarus(2011)	1,506	2.51	2.47	2.56	57	2.17	3.26	3.41	56	3.34	2.13	2.21	56
9	Bolivia(2017)	1,965	1.83	1.80	1.87	92	2.13	3.19	3.33	64	3.26	2.09	2.17	64
10	Brazil(2018)	1,557	2.46	2.42	2.51	66	2.40	3.71	3.87	31	3.80	2.36	2.44	32
11	Bulgaria(2006)	898	2.53	2.46	2.58	55	2.16	3.20	3.43	57	3.33	2.10	2.21	57
12	Burkina Faso(2007)	736	2.40	2.32	2.48	74	2.09	3.07	3.31	68	3.18	2.04	2.15	68
13	Canada(2020)	3,983	2.80	2.78	2.83	32	2.29	3.53	3.63	42	3.59	2.26	2.32	42
14	Chile(2018)	894	2.40	2.34	2.45	75	2.14	3.16	3.38	63	3.28	2.08	2.19	63
15	China(2018)	2,771	3.12	3.09	3.15	10	2.65	4.24	4.36	11	4.31	2.62	2.68	11
16	Colombia(2018)	1,440	2.16	2.11	2.22	83	2.22	3.34	3.52	49	3.43	2.17	2.26	49
17	Cyprus(2019)	903	2.67	2.60	2.75	43	2.35	3.60	3.84	38	3.71	2.30	2.42	38
18	Czechia(2022)	1,182	2.81	2.76	2.84	30	2.37	3.66	3.82	36	3.73	2.33	2.41	36
19	Ecuador(2018)	1,168	2.55	2.50	2.62	53	2.64	4.19	4.38	12	4.27	2.60	2.69	12
20	Egypt(2013)	1,056	2.44	2.43	2.55	70	1.87	2.73	2.89	89	2.74	1.87	1.95	89
21	Estonia(2011)	1,447	2.93	2.89	2.97	21	2.46	3.85	4.02	26	3.92	2.42	2.51	26
22	Ethiopia(2020)	916	2.53	2.47	2.60	56	1.98	2.85	3.08	82	2.96	1.93	2.04	82
23	Finland(2005)	972	3.26	3.22	3.30	5	2.49	3.88	4.07	23	3.99	2.43	2.53	23
24	France(2006)	964	2.75	2.70	2.79	38	2.17	3.24	3.43	55	3.34	2.12	2.22	55
25	Georgia(2014)	1,125	2.46	2.40	2.50	68	2.01	2.91	3.09	80	3.01	1.96	2.04	80
26	Germany(2018)	1,418	3.09	3.07	3.14	14	2.55	4.05	4.20	17	4.11	2.52	2.60	17
27	Ghana(2012)	1,444	2.63	2.57	2.68	47	2.54	3.97	4.16	18	4.07	2.49	2.58	18
28	Greece(2017)	1,132	2.82	2.76	2.86	28	2.50	3.90	4.08	21	4.00	2.45	2.54	21
29	Guatemala(2020)	1,011	1.84	1.80	1.89	91	1.85	2.60	2.80	92	2.69	1.80	1.90	91
30	Haiti(2016)	1,747	1.65	1.62	1.67	95	1.85	2.61	2.73	91	2.67	1.82	1.88	92
31	Hong Kong(2018)	1,994	2.69	2.65	2.72	40	2.35	3.61	3.75	40	3.69	2.31	2.37	40
32	Hungary(2009)	970	2.40	2.34	2.45	76	1.90	2.70	2.90	88	2.79	1.85	1.95	88
33	India(2023)	1,476	2.89	2.84	2.96	23	2.41	3.74	3.94	30	3.82	2.37	2.47	30
34	Indonesia(2018)	2,879	2.86	2.84	2.91	26	2.67	4.30	4.41	10	4.35	2.65	2.71	10
35	Iran(2020)	1,423	3.26	3.22	3.31	4	2.78	4.47	4.66	4	4.55	2.73	2.83	4
36	Iraq(2018)	1,024	2.64	2.58	2.73	45	2.21	3.33	3.54	50	3.42	2.17	2.27	50
37	Italy(2005)	932	2.91	2.87	2.96	22	2.49	3.91	4.10	22	4.00	2.45	2.54	22
38	Japan(2019)	1,013	2.96	2.93	3.01	19	2.71	4.35	4.52	5	4.43	2.67	2.75	5
39	Jordan(2018)	1,130	3.63	3.59	3.67	2	3.06	5.01	5.19	1	5.11	3.01	3.09	1
40	Kazakhstan(2018)	1,149	2.80	2.75	2.85	33	2.40	3.72	3.90	32	3.80	2.36	2.45	31
41	Kenya(2021)	1,171	2.20	2.15	2.26	81	2.03	2.96	3.14	77	3.06	1.98	2.07	78
42	Kuwait(2014)	1,105	3.03	2.95	3.07	16	2.48	3.83	4.04	24	3.96	2.42	2.52	24
43	Kyrgyzstan(2020)	1,153	2.42	2.37	2.48	72	2.07	3.04	3.23	73	3.13	2.02	2.12	73
44	Lebanon(2018)	1,175	2.67	2.61	2.72	42	2.35	3.58	3.80	39	3.70	2.29	2.40	39
45	Libya(2022)	1,146	2.60	2.52	2.64	48	2.15	3.18	3.39	61	3.30	2.09	2.19	61
46	Macao(2019)	780	2.87	2.82	2.93	25	2.39	3.70	3.91	33	3.79	2.34	2.45	33
47	Malaysia(2018)	1,291	2.68	2.64	2.73	41	2.37	3.65	3.81	35	3.74	2.33	2.41	35

48	Maldives(2021)	894	2.75	2.70	2.81	37	2.47	3.85	4.04	25	3.94	2.42	2.52	25
49	Mali(2007)	596	2.80	2.74	2.91	31	2.26	3.38	3.70	43	3.51	2.19	2.35	43
50	Mexico(2018)	1,655	1.81	1.77	1.86	93	1.94	2.78	2.95	85	2.87	1.89	1.98	86
51	Moldova(2006)	1,008	1.97	1.92	2.04	88	1.96	2.85	3.03	84	2.91	1.93	2.02	84
52	Mongolia(2020)	1,572	2.51	2.47	2.56	58	2.11	3.13	3.31	66	3.22	2.07	2.16	66
53	Morocco(2021)	842	2.47	2.39	2.53	65	2.25	3.37	3.61	44	3.50	2.19	2.31	44
54	Myanmar(2020)	1,157	2.58	2.52	2.63	50	2.16	3.21	3.41	58	3.31	2.12	2.21	60
55	Netherlands(2022)	1,689	2.79	2.76	2.83	35	1.93	2.81	2.95	86	2.87	1.90	1.97	85
56	New Zealand(2020)	846	3.21	3.17	3.27	6	2.58	4.07	4.30	14	4.17	2.53	2.65	14
57	Nicaragua(2020)	1,131	2.05	1.99	2.12	87	2.24	3.39	3.57	46	3.47	2.20	2.29	46
58	Nigeria(2018)	1,002	1.92	1.86	2.00	90	1.86	2.59	2.79	90	2.70	1.81	1.91	90
59	N. Ireland(2022)	385	2.73	2.64	2.80	39	2.09	3.01	3.34	69	3.18	2.01	2.17	69
60	Norway(2007)	999	3.05	3.01	3.10	15	2.13	3.16	3.37	65	3.25	2.08	2.18	65
61	Pakistan(2018)	1,328	2.24	2.19	2.32	80	2.00	2.93	3.12	81	3.00	1.97	2.06	81
62	Palestine(2013)	890	2.55	2.50	2.63	54	2.23	3.36	3.57	48	3.47	2.18	2.28	47
63	Peru(2018)	1,282	1.95	1.90	2.00	89	2.18	3.27	3.45	53	3.35	2.13	2.23	54
64	Philippines(2019)	1,074	3.11	3.07	3.15	11	2.68	4.28	4.44	8	4.36	2.64	2.71	8
65	Poland(2012)	838	2.49	2.43	2.55	61	2.11	3.07	3.31	67	3.21	2.04	2.15	67
66	Puerto Rico(2018)	1,076	2.49	2.44	2.56	60	2.24	3.39	3.62	45	3.49	2.20	2.31	45
67	Qatar(2010)	988	3.71	3.68	3.75	1	3.05	5.00	5.17	2	5.10	3.00	3.08	2
68	Romania(2018)	1,075	2.47	2.40	2.53	63	2.38	3.64	3.86	34	3.75	2.32	2.43	34
69	Russia(2017)	1,668	2.47	2.41	2.50	64	2.16	3.22	3.37	59	3.32	2.11	2.19	58
70	Rwanda(2012)	1,321	2.76	2.71	2.79	36	2.09	3.07	3.26	70	3.17	2.04	2.14	70
71	Serbia(2017)	957	2.27	2.20	2.31	79	1.97	2.84	3.05	83	2.95	1.92	2.03	83
72	Singapore(2020)	1,862	3.10	3.07	3.13	13	2.70	4.33	4.46	6	4.40	2.66	2.73	6
73	Slovakia(2022)	1,183	2.46	2.40	2.50	69	2.04	2.96	3.15	76	3.07	1.98	2.08	76
74	Slovenia(2011)	1,009	2.37	2.32	2.42	77	2.15	3.22	3.40	60	3.31	2.11	2.20	59
75	South Africa(2013)	3,243	2.42	2.40	2.47	73	2.05	3.04	3.17	74	3.09	2.03	2.09	74
76	South Korea(2018)	1,241	2.57	2.54	2.61	51	2.18	3.27	3.43	52	3.35	2.14	2.21	53
77	Spain(2011)	1,087	2.59	2.53	2.64	49	2.03	2.99	3.18	78	3.07	2.00	2.09	77
78	Sweden(2011)	1,082	2.93	2.89	2.98	20	2.18	3.26	3.45	54	3.35	2.12	2.22	52
79	Switzerland(2007)	1,171	2.99	2.94	3.01	18	2.36	3.62	3.77	37	3.72	2.31	2.38	37
80	Taiwan(2019)	1,194	2.99	2.95	3.03	17	2.52	3.94	4.11	20	4.04	2.47	2.55	20
81	Tajikistan(2020)	1,145	3.20	3.18	3.28	7	2.70	4.36	4.52	7	4.39	2.68	2.76	7
82	Thailand(2018)	1,320	2.56	2.54	2.63	52	2.07	3.07	3.25	72	3.14	2.04	2.13	72
83	Trinidad(2010)	909	2.16	2.11	2.23	84	2.09	3.08	3.30	71	3.17	2.05	2.15	71
84	Tunisia(2019)	1,037	2.46	2.39	2.50	67	2.19	3.25	3.45	51	3.38	2.12	2.22	51
85	Turkey(2018)	2,100	3.18	3.14	3.21	8	2.67	4.27	4.41	9	4.35	2.63	2.71	9
86	Ukraine(2020)	1,087	2.20	2.16	2.27	82	1.91	2.71	2.92	87	2.80	1.86	1.97	87
87	UK(2022)	2,200	2.80	2.76	2.83	34	2.14	3.20	3.35	62	3.29	2.10	2.17	62
88	United States(2017)	2,517	2.82	2.78	2.85	29	2.45	3.83	3.96	28	3.90	2.42	2.48	28
89	Uruguay(2022)	944	2.82	2.73	2.86	27	2.56	3.91	4.18	16	4.12	2.46	2.59	16
90	Uzbekistan(2022)	1,145	3.27	3.22	3.31	3	2.85	4.58	4.76	3	4.69	2.79	2.88	3
91	Venezuela(2021)	1,179	1.72	1.67	1.76	94	1.77	2.46	2.60	93	2.53	1.73	1.81	93
92	Vietnam(2020)	1,188	3.16	3.12	3.20	9	2.57	4.04	4.21	15	4.14	2.52	2.60	15
93	Yemen(2014)	569	2.11	2.06	2.21	85	1.73	2.40	2.62	95	2.47	1.70	1.81	95
94	Zambia(2007)	1,296	2.50	2.45	2.55	59	2.32	3.54	3.73	41	3.64	2.27	2.36	41
95	Zimbabwe(2020)	1,143	2.48	2.42	2.54	62	2.23	3.36	3.56	47	3.47	2.18	2.28	48

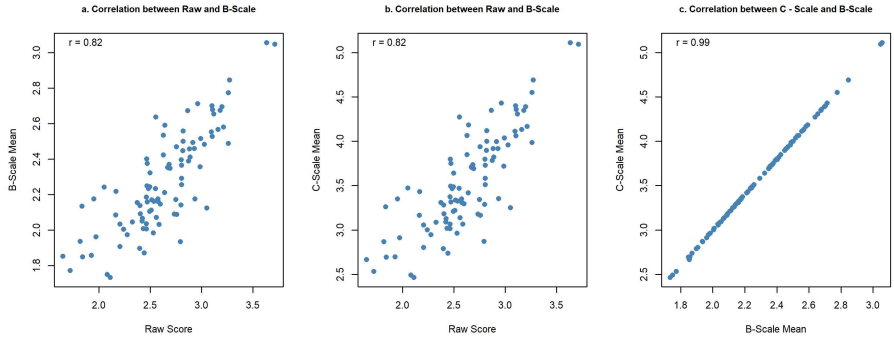
원 자료: WVS Time Series 1981~2022 Data

주 1) 고정항목의 동점(Ties) 처리는 "균등배당(Uinform Allocation)" 방식 적용

주 2) 보정 점수의 통계적 신뢰도와 순위 변동의 유의성을 검증하기 위해 비모수적 부트스트랩 기법 (200회 재복원추출)을 적용하여 95% 신뢰구간의 하한과 상한을 도출함

[부록]

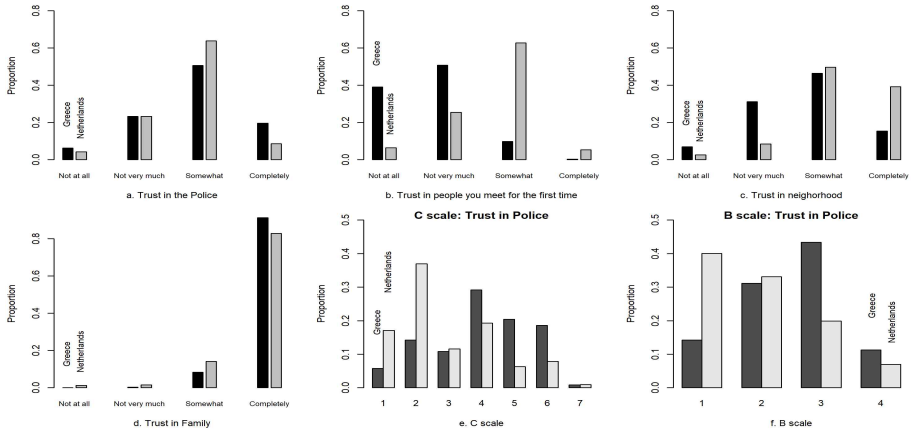
〈그림 1〉 WVS 국가별 경찰신뢰의 원점수, C Scale, B Scale 간 상관관계



원 자료: WVS Time Series 1981~2022 Data

주 1: 고정항목의 동점(Ties) 처리는 "균등배당(Uinform Allocation)" 방식 적용

〈그림 2〉 그리스와 네덜란드의 경찰신뢰에 대한 비모수적 측정 적용



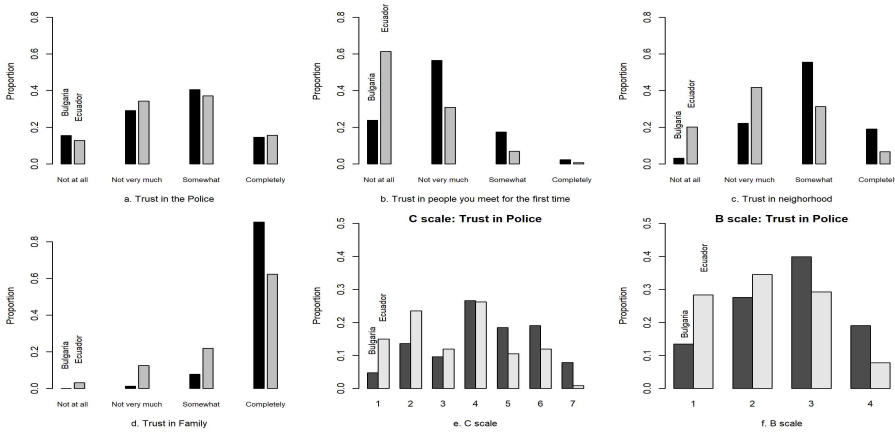
원 자료: WVS Time Series 1981~2022 Data

주 1: 그리스 경찰신뢰(N=1,200, year = 2017, 원자료 평균 = 2.82, C scale 평균 = 4.00, B scale 평균 = 2.50), 네덜란드 경찰신뢰(N=2,145, year = 2022, 원자료 평균 =2.79, C scale 평균 = 2.87, B scale 평균 = 1.93)

주 2: 고정항목의 동점(Ties) 처리는 "균등배당(Uinform Allocation)" 방식 적용

[부록]

〈그림 3〉 불가리아와 에콰도르의 경찰신뢰에 대한 비모수적 측정 적용

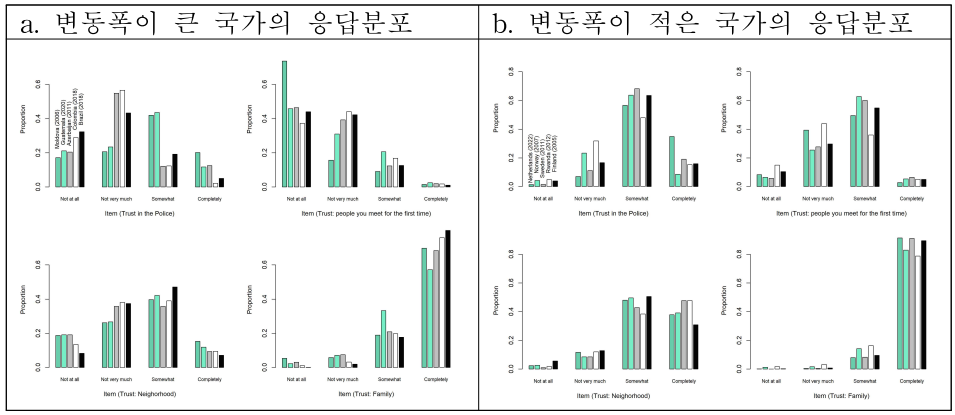


원 자료: WVS Time Series 1981~2022 Data

주 1: 불가리아 경찰신뢰(N=1,001, year = 2006, 원자료 평균 = 2.53 C scale 평균 = 3.33 , B scale 평균 = 2.16), 에콰도르 경찰신뢰(N=1,200, year = 2018 원자료 평균 = 2.55, C scale 평균 = 4.27, B scale 평균 = 2.64)

주 2: 고정항목의 동점(Ties) 처리는 "균등배당(Uinform Allocation)" 방식 적용

〈그림 4〉 보정 전후 변동폭이 큰 국가와 적은 국가의 응답분포



원 자료: WVS Time Series 1981~2022 Data

주 1: 변동폭이 큰 국가는 네덜란드(2022년), 노르웨이(2007년), 스웨덴(2011년), 르완다(2012년), 핀란드(2005년)이고, 적은 국가는 몰도바(2006년), 과테말라(2020년), 아제르바이잔(2011년), 콜롬비아(2018년), 브라질(2018년)임

[부록]

〈표 2〉 분석 수준 및 척도별 CPI와의 상관관계 비교

비교 대상	원점수 vs CPI	C 변환 vs CPI	B 변환 vs CPI
상관계수 (spearman's ρ)	0.467 ***	0.251 **	0.255 **

원 자료: WVS Time Series 1981~2022 Data, Transparency International (2025)
주 1) p-value * 〈 0.05, ** 〈 0.01, *** 0.001
주 2) 분석 대상 표본 수(N): 95개 국가에 대한 분석으로 원안 및 민감도 분석 조건 1~3은 결측치를 제외한 전체 응답자 수(약 212,129명)를 유지하였으나, 동점 처리 변경의 경우 서열 판단이 불가능한 동점 사례가 제외되어 유효 표본 수가 약 60,133명으로 감소함
주 3) 국가 간 순위 변동폭은 분석조건별 국가들의 순위(Rank)를 매긴 후, 그 순위의 차이(절대값)를 계산하여 산술평균함
주 4) 동점처리는 (1) 주어진 백터구간 내에서 균등하게 배분하는 방법에서 (2) 고정항목 동점사례를 삭제하는 방법으로 변경한 결과임

Abstract

A Study on Measuring Police Trust Cross-Nationally Using Anchoring Items

*

Building upon the problematic issues raised by King et al. (2004), this study aims to present a calibration method for measuring police trust using anchoring items and to empirically verify its validity, specifically to address Differential Item Functioning (DIF) —a critical measurement error in cross-national comparative research. Utilizing data from the World Values Survey (WVS) international survey project (1981 - 2022), this research proposes a non-parametric calibration methodology that repurposes existing trust items (family, neighbors, and strangers) as anchoring items. The analysis was conducted on 95 countries worldwide, with confidence intervals calculated using non-parametric bootstrap techniques to ensure statistical robustness. The results confirm that comparisons based on raw scores can result in inverted or distorted trust levels due to country-specific response tendencies, as evidenced by the cases of the United Kingdom and Japan. In particular, the post-calibration rank shifts in key countries such as Finland, Nicaragua, and the Netherlands demonstrated statistical significance that transcends mere sampling error. This study holds both academic and policy significance by highlighting the necessity of controlling for DIF in international comparative research within the fields of public policy and administration, while providing a balanced presentation of the procedures, benefits, and limitations of using anchoring items for measurement calibration.

Key words : differential item functioning, police trust, cross-national comparison, anchoring items, World Value Survey