검사 번역/번안을 위한 국제 지침: 한국어판
2017 년 2 판

번역자: 서동기(한림대학교 심리학과 교수),
이순묵(성균관대학교 심리학과 명예교수)

Reference this document as:
International Test Commission. (2017). The ITC Guidelines for Translating and Adapting Tests (Second edition). [www.InTestCom.org] *Translation authorized by the Korean Psychological Association.*

Translated by: Dong Gi Seo (Hallym University),
Soonmook Lee (Sungkyunkwan University)

검사 번역 및 번안을 위한 ITC 지침서

(2판)

역자서문: 원본 문헌을 또 다른 문화권으로 옮겨가기 위해서는 언어적 의미뿐만 아니라 원본 문화권과 대상 문화권 사이의 심리적/문화적 차이를 고려하는 노력이 필요함을 역자들은 인식하였다. 따라서 본 지침서의 번역에서 단순히 언어적 의미를 넘어 영어권과 한국어권에서의 심리적/문화적 차이를 고려하는 번역을 위해 본문에서 [역자주]를 간간이 삽입하였고 독자의 이해를 돕고자 원본 지침서에 없는 96개의 각주를 제공하였다. 그리고 이 지침서에서 측정은, 경험되는 사실들에 숫자를 부여하는 작업이고 검사는 행동의 표본에 대한 측정을 가능하게 하는 모든 종류의 논리, 방법, 절차, 도구를 의미한다. 따라서 필기시험, 실기시험, 설문, 심리검사, 능력 수준에 대한 등급매기기, 또는 합격/불합격 판정하기 등이 모두 검사에 포함된다.

원본: International Test Commission. (2017). ITC Guidelines for Translating and Adapting Tests (Second Edition), International Journal of Testing, 18:2, 101-134, DOI: 10.1080/15305058.2017.1398166 [1)]

검사 번역 및 번안을 위한 국제검사위원회 (ITC) 지침서 제2판은 초판을 보완하고 검사 기술과 시행의 발전을 반영하기 위하여 2005년부터 2015년까지 준비되었다. 지침서에는 여섯가지 범주로 나누어지는 18가지 지침이 존재한다: 선행조건(pre-condition), 검사개발(test development), 검수(confirmation), 실시(administration), 점수부여 및 결과해석(scoring and interpretation), 문서화(documentation). 각 지침에 대한 설명 및 실제 적용에 대한 안내를 제공한다. 또한 지침의 원활한 이행을 위해 점검표를 제공한다.

핵심어: 검사 번역, 검사 번안, 검사의 현지화, 측정의 동일성, 구성개념 동등성

지난 25년 동안 검사의 번역 및 번안 분야의 방법론은 몇몇 책의 출판과 많은 연구들, 그리고 뛰어난 검사 번안 작업의 사례 등을 통해 급속히 발전하였다(Grégoire, & Hambleton, 2009;

[1)] https://doi.org/10.1080/15305058.2017.1398166

Hambleton, Merenda, & Spielberger, 2005; Rios, & Sireci, 2014; van de Vijver, & Leung, 1997, 2000). 이러한 발전은 다음 분야들에서 검사의 번역 및 번안에 대한 관심이 증가하면서 필연적이게 되었다.

(1) 비교 문화 심리학

(2) 교육 성취에 대한 대규모의 국제적인 비교 연구 (예, TIMSS, OECD/PISA)

(3) 국제적으로 통용되는 자격 인증 시험 (예, Microsoft나 Cisco와 같은 정보 기술 분야)

(4) 응답자가 검사 수행시 언어를 선택할 수 있도록 허용함으로써 검사의 공정성 제고 (이스라엘의 경우 대학 입학을 위해 여섯개 언어 중 하나를 선택하여 시험에 응시할 수 있다).

번안된 검사(test)와 설문지(questionnaires)의 구성개념(construct), 방법(method), 문항 편향(bias)의 평가를 위한 질적이고 양적인 접근에서 기술적 진보가 이루어졌는데, 여기에는 문항반응이론(Item Response Theory; IRT), 구조방정식모형(Structural Equation Model; SEM), 일반화가능도이론(Generalizability theory)과 같은 복잡한 통계적 절차의 사용이 포함된다(Hambleton 등, 2005; Byrne, 2008). 또한 OECD/PISA에 의해 새로운 번역 방법이 발전되었고(Grisay, 2003) 검사 번안 프로젝트를 완성해 가는 단계가 제시되었다(Hambleton, & Patsula, 1999). 검사번안 실무를 안내하는데 전형이 되는 프로젝트들 가운데 OECD/PISA 와 TIMSS 프로젝트가 있다. 이외에도 많은 사례에서 발전이 이루어 졌다.

지침서 초판은(Hambleton, 2005; van de Vijver & Hambleton, 1996) 응답자 집단 간 비교를 가능하게 하고 원할하게 하는 검사 번안을 목적으로 하는 비교연구의 관점에서 출발하였다. 지침서 작성에 내재된 기본 틀은, 하나의 문화에서 개발된 검사는 다른 문화의 맥락에 맞는 검사 개발로 이어진다는 것이다(기존의 도구는 새로운 문화적 맥락에서 사용할 수 있도록 번안되어야 한다). 하지만 검사 번안이 더욱 더 넓은 적용 분야를 포함하고 있음이 명확해지고 있다. 가장 중요한 예는 다문화 집단에서 기존의 검사를 사용할 것인가 아니면 그 집단에 맞는 새로운 검사를 개발할 것인가이다. 이를테면, 상담장면에서 서로 다른 인종 집단의 내담자들, 검사 언어의 숙달 정도가 다른 다양한 인종 집단에 대한 교육평가, 다국적 기업의 경영을 위한 국제적 채용을 들 수 있다. 이렇게 검사 번안이 가능한 영역의 확장은 검사의 개발, 실시, 타당화 및 문서화에 영향을 미친다. 가능한 예로서, 기존 검사 문항에 대한 비원어민의 이해도 향상을 위해 번안이 필요하다는 것이다(예를 들어 언어의 단순화). 지침서의 또 다른 중요한 확장은 동시개발(즉, 원언어와 현지언어로 된 도구의 동시개발)을 수용하는 것이다. 국제적인 대규모 검사 프로젝트의 경우 한 언어로 개발된 검사를 나중에 해당 연구에 관련된 모든 언어로 번역하거나 번안할 수 없게 되는 상황을 피하고자 동시개발 방법을 사용하고 있다.

검사 번역 및 번안에 대한 ITC 지침의 초판은 Vijver와 Hambleton(1996), Hambleton(2002), Hambleton 과 동료들(2005)에 의해 출판되었다. 1996년에서 2005년 사이에

출판된 지침서 들에서는 편집상의 사소한 수정만이 있었다. 한편 1996년 이래로 많은 발전이 있었다. 첫째, ITC 지침에 대한 유용한 검토가 많이 제시되었는데, Tanzer 와 Sim(1999), Jeanrie 와 Bertrand(1999), 그리고 Hambleton(2002)의 논문이 여기 포함된다. 이들 저자들은 지침이 가지고 있는 가치를 강조하며 개선을 위한 일련의 제안을 제시했다. Hambleton과 동료들(2005)은 1999년 미국 조지 타운 대학에서 개최된 ITC 국제학회에서 논문집을 출판했는데 그중 몇 개 장의 저자들 즉, Cook 과 Schmitt-Cascallar(2005), 그리고 Sireci(2005)는 검사 번안을 위한 새로운 패러다임을 발전시키고 새로운 방법론을 제시하였다. ITC는 2006년 벨기에 브뤼셀에서 검사 번역 및 번안에 대한 ITC 지침에 초점을 맞춘 국제 학회를 개최했다. 40개국 이상에서 400명 이상의 사람들이 검사 번안이라는 주제에 중점을 두고 새로운 방법론을 내놓았고, 새로운 지침을 제안하였으며, 성공적인 구현 사례를 공유하였다. 1996년에서 2009년까지 국제 학회의 심포지엄에서 많은 논문이 발표 되었고(예, Grégoire & Hambleton, 2009) ITC 지침서 제2판의 초기 스페인어 판을 Muñiz, Elosua, and Hambleton(2013)의 논문에서 볼 수 있다.

2007년, ITC 이사회는 6인의 위원회를 결성해 이 분야에서 진전된 새로운 지식과 연구자들이 얻은 많은 경험을 강조하여 ITC 지침을 갱신하도록 하였다. 갱신된 내용은 다음과 같다.

(1) 서로 다른 언어 집단 간 검사의 요인 동등성 확인을 위한 구조방정식모형의 발전
(2) 서로 다른 언어 집단 간 다분문항 평정척도(rating scale)에서의 차별기능문항 식별을 위한 확장된 접근
(3) OECD/PISA 및 TIMSS와 같은 국제적인 평가 프로젝트에 의해 개척된 새로운 번안 설계

이 위원회 역시 2008년 프라하, 2009년 오슬로에서 개최된 심리학자들의 국제 학회에서 새로운 지침에 대한 발표와 초안을 제공하고 상당한 정도의 피드백을 받았다.

실시 지침에 대한 절은 제2판에서도 유지되었으나 중복된 지침은 결합되어, 전체적으로 지침의 수가 여섯 개에서 두 개로 축소되었다. "문서화/점수해석"은 초판에서는 마지막 절이었다. 제2판에서 이 절은 독립된 두 개의 절- 점수부여와 해석, 문서화로 나누었다. 또한 이 절에 원래 들어 있던 네 개 지침 중 두개는 상당히 개정되었다.

초판과 마찬가지로, 우리는 검사 번역과 번안의 차이점을 독자들이 분명하게 구분하기를 원한다. 검사 번역(translation)이 아마도 일반적인 용어이지만, 번안(adaptation)은 더 넓은 의미의 용어이며 검사를 한 언어와 문화에서 다른 언어와 문화로 옮겨가는 것을 가리킨다. 검사 번안은 제 2의 언어와 문화에서 번안된 검사가 원본언어에서와 같은 구성 개념을 측정하는지 아닌지의 결정, 번역자를 선정하는 일, 검사 번역을 평가하기 위한 설계(design)의 선택(예, 번역-역번역),

필요한 모든 양해사항(accommodation)[2]의 선택, 검사 형식의 수정, 번역의 실시, 제 2의 언어와 문화에서 수행된 검사의 동등성을 확인, 기타 필요한 타당화 연구의 진행을 포함한 모든 활동을 말한다. 반면, 검사 번역은 한 언어와 문화에서 다른 언어와 문화로 검사를 옮겨갈 때 언어적 의미를 보존하기 위한 언어의 실제 선택이라고 하는 제한된 의미를 지니고 있다. 검사 번역은 번안의 일부일 뿐이지만, 그 자체로만 보자면 한 언어와 문화에서 다른 언어와 문화로 검사를 옮겨감에 있어서 교육학적 혹은 심리학적 동등성에 대한 고려가 없는 매우 단순한 접근이다.

지침

서론

이 지침서에서의 지침을 정의하자면, [원본검사 집단이 아닌] 다른 집단에서의 사용을 목적으로 한 심리학적/교육학적 검사의 동시개발을 수행하고 평가하는데 중요한 실무를 의미한다. 지침은 여섯개 주제로 구성되었고 총 18개이다: 선행조건(3개 지침), 검사개발(5개 지침), 검수-경험적 분석(4개 지침), 실시(2개 지침), 점수부여및 결과해석(2개 지침), 문서화(2개 지침).

제 1절 인 "선행조건" 지침은 번역/번안 과정이 시작되기 전에 필요한 결정이 내려져야 한다는 사실을 강조한다. 제 2절인 "검사개발" 지침은 검사 번안의 실제 과정에 중점을 두었다. 제 3절인 "검수" 지침은 다양한 언어와 문화에서 검사의 동등화[3], 신뢰도, 타당도를 지지하는 경험적 증거의 누적과 연관된 것이다. 마지막 세 절은 "실시", "점수부여 및 결과해석", "문서화"와 관련되어 있다. "문서화"는 심리학 및 교육학의 번안 노력에서 특히 등한시 되어오던 주제였다. 우리는 학회지 편집자 및 연구지원기관에서 번안 과정의 문서화에 대해서 더욱 엄격히 고려할 것을 기대한다. 각 지침에서는 설명 및 실무에서의 시행을 위한 제안을 제시하고 있다.

1. 선행조건 지침

선행조건-1 (1)[4]. 번안하기 전에 검사와 관련된 지적 재산권 소유자로부터 허가를 얻는다.

---

[2] 양해사항은 검사 응답자 개인에게 특별히 다른 사람들에 비해 불리한 점을 감안하여 검사환경가운데 필요한 정도로 배려가 주어지는 부분들임. 자세한 정의는 아래 웹사이트에서 확인가능하다. https://www.edglossary.org/test-accommodations/

[3] 원본 검사를 기준으로 하는 동등화를 말한다.

[4] 선행조건-1은 "선행조건" 관련의 지침 3개중 1번임을 의미하고 (1)은 전체 18개 지침중에서

설명. 지적재산권(Intellectual property rights)은 자신의 창작물, 발명품, 또는 제품에 대해 가지는 일련의 권리를 의미한다. 이는 창작물에 대한 도덕적, 경제적 권리를 부여함으로써 창작자의 이익을 보호한다. 세계지적재산권기구(www.wipo.int)에 따르면, "지적재산은 정보나 지식의 항목에 대한 것으로서, 이는 어느 곳에서든 동시에 무제한 복제되어 유형물로 나타날 수 있는 것들이다."

지적재산에는 산업재산과 저작권이 있다. 산업재산은 발명품, 산업디자인, 상표 및 상호를 보호하는 특허를 가리킨다. 저작권은 예술적 혹은 기술 기반의 창작물을 의미한다. 창작자(저자)는 자신의 창작물에 대해 특정의 권한을 가진다(예, 복제되거나 각색될 때 왜곡 방지). 그외 사람들(예, 출판사)은 창작자(저자)로부터 허가 받은 경우에 그에 따른 권리를 행사할 수 있다(예, 복사본 만들기). 많은 경우 검사에 대한 저작권은 다른 저술과 마찬가지로 저자가 출판사 혹은 배포자에게 양도할 수 있다[5].

교육학적/심리학적 검사는 분명히 인간 정신의 창작물[6]이기 때문에 지적재산권에 의해 보호된다. 대부분 저작권은 문항의 특정 내용(예, "1+1=...?" 혹은 "나는 슬프다" 등과 같은 문항에 대해서는 저작권이 없음)이 아닌 검사의 독창적인 구성(척도의 구조, 점수부여, 자료의 조직화 등)에 해당된다. 따라서, 기존 검사의 구조와 점수부여 체계를 그대로 유지하면서 새로운 문항을 만드는 것은 모방(mimicking)으로 지적재산권에 대한 침해이다. 번안에 대한 권한을 인가받은 경우, 검사 개발자는 원저자로부터 기존 검사의 특성(구조, 내용, 형식, 점수부여 등)을 수정할 수 있도록 허락받지 않은 이상 이를 존중해야 한다.

실무를 위한 제안. 검사 개발자는 기존 검사의 모든 저작권법 및 계약을 존중해야 한다. 번안 작업을 시작하기에 앞서 지적재산권 소유자(저자 또는 출판사)가 서명한 계약서를 가지고 있어야 한다. 계약서에는 기존 검사의 특성과 관련하여 수정 가능한 사항을 명시해야 하며 번안된 검사에 대한 지적재산권 소유자를 명확히 해야 한다.

선행조건-2 (2). 검사점수를 적용하고자 하는 대상 집단에 비추어, 검사에서 측정될 구성개념의 정의와 내용이 문항내용에 반영된 정도가 충분한지 평가한다[7].

1번임을 의미한다.

[5] 한국내에서는 저작권은 저자에게 있고 출판사는 판권을 가진다.

[6] 예술적인 것은 아니지만 과학적 지식과 기술을 기반으로 하는 창작물이다.

[7] 문항의 '구성개념 표상(construct representation)'이라고 하며 검사의 구성개념 타당도에 대한

설명. 이 지침에서는, 검사에서 무엇이 평가되는지에 대하여 언어 및 문화가 다른 집단 간에 동일하게 이해될 것을 요구하며, 이것이 문화 간 타당한 비교의 기초가 된다. 이 단계에서 검사나 도구가 아직은 번안되지 않았으므로, 연구대상 집단 내에서 기존의 유사한 검사들에서의 경험적 증거를 살펴보거나, 구성개념과 문항 간 대응에 대한 판단, 그리고 연구대상 집단의 언어에 적절한 검사인지를 판단하는 것이 바람직하다고 할 수 있다. 그러나 궁극적으로 이 중요한 지침은, 검수-2 (10)에서 요구하는, 증거에 수반되는 경험 자료[8]에 비추어 평가되어야 한다. 분석의 목적은 검사의 구조를 확립하는 것이 아니라(그 확립이 분석의 부산물이긴 하지만), 여러 언어판 검사에 걸친 구조의 동등성을 검수하는 것이다.

실무를 위한 제안. 각 문화/언어 집단에서 측정되는 구성개념의 적절성을 평가하기 위해서는, 측정되는 구성개념과 관련된 전문가들 그리고 평가대상인 문화집단에 친숙한 전문가들을 모집해야 한다. 그리고 모집된 피험자들은 다음 질문에 답을 해야 할 것이다. 즉, 구성개념이 두 집단(원검사 집단, 새로운 문화집단)의 문화들에서 이해 가능한 것인지? 예를 들어, 위원회(ITC)에서는 [원래 문화의]교육장면검사에서 측정된 구성개념이 제2의 문화에서는 의미가 없거나 퇴색되는 것으로 판단된 경우가 여러 번 있었다. 초점 집단, 면담 및 설문 조사와 같은 방법을 사용하여 구성개념의 동등성 정도에 대한 구조화된 정보를 얻을 수 있다.

선행조건-3 (3). 대상집단[9]에서 검사 의도와 관련 없는 언어/문화적 차이가 가져오는 영향을 최소화한다.

설명. 검사가 측정하고자 하는 변수[10]와 무관한 문화적, 언어적 특성은 [번안]프로젝트의

증거 중 하나가 된다.

[8] 여기서 증거는 '집단 간에 구성개념의 동등성, [측정]방법 동등성 및 문항 동등성' 증거이고, 경험자료는 그 증거들에 대한 관련 통계수치들이다. 선행조건-2(2)에서는 구성개념이 문항에 잘 표상되고 있는지의 질적평가를 필요로 하고 검수-2(10)에서는 집단 간에 구성개념이 동등한지 양적평가를 필요로 하는 것이다.

[9] 이 지침서 전체에서 "population"은 검사대상 집단을 가리킨다. 따라서 특별히 "검사"라는 접두사 없이 대상집단으로 번역한다.

[10] 여기서 변수는 구성개념의 의미이다. 구성개념이 전통적으로는 심리학에서 이야기하는 잠재변수(latent variable, factor)이지만, 검사의 관점에서는 측정의 대상이 되는 구체적 행동(예, 구매시도, 범죄가담정도 등)이나 복합변수(composite variable, component, 예, 삶의 질을 "경제+건강+주관적 안녕감"으로 측정할 경우)도 포함된다.

초기 단계에서 식별 되어야한다. 그런 무관한 특성들 가운데는 검사에서 사용되는 문항의 형식, 자료 (예, 컴퓨터, 그림, 도표의 사용 등), 시간 제한 등이 있다.

이 문제에 대한 접근은 원검사 집단과 언어적/문화적으로 다른 대상집단 간의 '언어적/문화적 거리'를 평가하는 것이다. 언어적/문화적 거리의 평가에는 언어, 가족 구조, 종교, 생활 방식 및 가치의 차이에 대한 고려가 포함될 수 있다 (van de Vijver & Leung, 1997).

이 지침은 주로 질적 방법론에 그리고 특정 문화 및 언어의 차이에 대한 연구에 익숙한 전문가들에게 의존한다. 검사 번역자는 대상언어와 문화에 대해 원어민이어야 하지만, 대상언어만 알면 방법 편향에 대한 가능한 원인들을 식별하기 어렵기 때문에, 번역자 선정에 대한 특별한 주의가 필요하다. 예를 들어, Hambleton, Yu 및 Slater (1999)에 의해 수행된 8학년 중국계 미국인의 수학 성취도에 대한 비교 연구에서는, 수학 시험관련의 많은 문화적 측면들과 더불어 형식 및 시험의 길이에서 문제가 발견되었다.

실무를 위한 제안. 이 지침은 어떤 시점에서도 경험적 자료를 구해서 논하기가 어렵다. 특히 번안 초기 단계에서는 경험적 자료에 근거한 논의를 하기 어렵다. [따라서] 초기단계에서는 질적 증거를 수집해야 하는 일이 자주 있다.

- 관찰, 면담, 초점 집단 또는 설문조사를 통해 검사 참가자의 동기 부여 수준, 지시사항들에 대한 이해, 심리검사 경험, 검사실시의 신속성, 평정 척도(rating scale)에 대한 친숙도 및 문화적 차이를 파악한다(단, 제반 변수[11] 자체에 대한 이해가 문화 간에 다르면 이러한 검토가 무의미해질 수 있다). 참가자들로부터 이러한 연구 자료를 수집하는 것에 문제가 있으면 번역자들로부터 가능한 많은 정보를 구해야 한다. 이들 작업의 일부는 검사 번안을 진행하기 전에 이루어질 수 있다.
- 일단 검사가 번안되고, 공분산분석이나 기타 분석들을 통한 타당화 연구의 준비가 되었으면, 이어지는 경험분석에서 이들 '오염변수'를 통제하는 것이 가능하다. 여기서 타당화 연구는 여러 언어/문화 집단으로 구분되는 참가자들을 동기부여 수준이나 특정의 평정척도에 대한 친숙도 수준에 맞추어 분류하고 분석하는 것이다[12](e.g., Javaras & Ripley, 2007; Johnson, 2003).

---

[11] 여기서 제반 변수는 '검사 참가자의 동기부여 수준, 지시사항에 대한 이해, 심리검사 경험, 검사실시의 신속성, 평정척도에 대한 친숙도 및 문화적 차이'를 가리킨다. 이들은 다음 문단에서 말하는 '오염변수'이다.

[12] 이러한 분석은 실험설계의 틀에서 보면 block factorial design이 될 수 있다. 즉 언어/문화 차이가 블록이 되고, 동기부여 및 친숙도 변수들이 요인이 된다.

2. 검사개발 지침

검사개발-1 (4). 번역 및 번안 과정에서 관련 전문 지식을 갖춘 전문가를 택하여, 대상 집단의 언어적, 심리적, 문화적 차이에 대한 고려를 확실히 한다.

설명. 이 지침은 검사 번안 기관이 검사 번안과 관련된 번역자를 구할 때, 두 언어에 대한 지식을 넘어 유자격자[13]를 찾도록 하는 데 영향을 미쳤음을 보여주는 상당한 증거로 인하여 수년에 걸쳐 가장 영향력 있는 지침 중 하나가 되었다(Grisay, 2003). 문화에 대한 지식 그리고 검사 주제와 제작에 대한 최소한의 일반적 지식은 번역자 선정 기준의 일부가 되었다. 또한, 이 지침은 다양한 설계(예, 번역 및 역번역 설계)에서 검사를 번역 및 번안하는 기관으로 하여금 최소한 두 명의 번역자를 사용하도록 장려하는 방향으로 영향을 주었다. 모든 결정 과정에서 한 명의 번역자에 의지하는 낡은 관행은 오늘날에는 수용 가능한 관행의 목록에서 제외되었다.

대상 문화에 대한 전문 지식은 번역자가 대상언어에 대한 원어민일 것을 필수로 하고, 대상언어권의 현장에서 생활하는 것을 바람직한 조건으로 할 때 확보된다. 대상언어의 원어민 번역자는 정확한 번역을 할 뿐만 아니라 부드럽게 읽히고 문화적으로 동질감을 주는 번역본을 만든다. 번역자가 대상언어권의 현장에서 살고 있다면 현재 사용되는 언어에 대한 최신 지식을 가지고 있을 것이 확실하다.

지침에서 '전문가'에 대한 정의는 전문적 품질의 번역/번안 검사의 제작에 필요한 (1) 관련 언어, (2) 문화, (3) 검사의 내용, 그리고 (4) 검사 전반(testing)[14]에 대한 일반적인 원칙에 대하여 충분한 지식을 가진 사람 또는 집단이다. 다른 사람들이 간과할 수 있는 영역을 검토하기 위해, 서로 다른 역량을 갖춘 사람들(예, 특정 주제에 전문 지식이 있는/없는 번역자, 검사 전문가 등)로 구성된 집단을 사용하는 것이 효과적일 수 있다. 모든 경우에 검사 내용에 대한 지식 외에도 검사 전반의 일반 원칙에 대한 지식은 번역자 훈련내용의 일부로써 제공되어야 한다.

실무를 위한 제안. 다음과 같은 세부 사항들을 제안한다.

- 대상언어권에서의 원어민이면서 그 문화에 대한 심층적 지식이 있는 사람을 번역자로 선정한다. 대상언어권 현지에 거주하는 사람이 선호된다. 흔히 범하는 실수는, 언어는

---

[13] 단순히 집단간 상이한 언어에 대한 지식만이 아닌 심리적/문화적 차이를 이해하는 번역자가 유자격자이다.

[14] 검사전반(testing)은 검사의 구상, 검사 제작 또는 번역/번안, 실시, 점수부여, 타당화, 문서화등의 제반 내용을 가리킨다

알고 있지만, 문화를 잘 알지 못하는 사람을 번역자로 선정하는 것이다. 문화에 대한 심층적 지식이 [검사의] 문화적 동질성[15]을 유지하는데 필수적인 경우가 자주 있다. 문화적 지식을 갖게 되면 현지 참가자들이 익숙하지 않은 문화 특수적 어휘들(예: 크리켓, 에펠 탑, 링컨 대통령, 캥거루 등)을 식별하게 된다.

- 가능하다면 검사 내용에 대한 경험과 평가 원리에 대한 지식을 가진 번역자를 선정한다 (예, 다지 선다형의 경우, 정답은 다른 선택지보다 길거나 짧지 않고 비슷한 정도라야 하며, 문법적 단서가 정답을 찾는 데 도움이 되지 않도록 하며, 참-거짓(진위형) 문항에서 참 진술이 거짓 진술보다 더 길지 않아야 한다).
- 검사 개발 원칙을 알고 있는 번역자를 찾는 것이 거의 불가능 할 수 있다. 따라서 번역자에게 자신이 작업할 형식에 따른 문항 작성의 원칙에 대하여 교육을 제공하는 것이 필수적이다. 훈련이 되지 않을 경우, 때로는 지나치게 양심적인 번역자는 번역된 검사의 타당도를 낮추는 오류의 원천이 될 수 있다. 예를 들어 번역자는 의도된 정답이 확실하게 실제 정답이 되도록, 의미를 분명히 하는 말을 추가할 수 있다. 그렇게 함으로써, 번역자는 그 문항을 의도한 것보다 더욱더 쉽게 만들 수 있고, 혹은 더 길어진 정답은 시험에 강한 참가자들에게 정답이라는 실마리를 제공할 수 있다.

검사개발-2 (5). 대상 집단에서 검사 번안의 적합성을 극대화하기 위해 적절한 번역 설계와 절차를 사용한다.

설명. 이 지침은 번역자 또는 번역 집단에서 내리는 결정이, 대상 집단에 대한 번안의 적합성을 최대한 높이는데 기여할 것을 요구한다. 이는 [번안검사에서의] 언어가 자연스럽고 수용할 만하다고 느껴져야 한다는 것을 의미한다. [원본 검사와 번안검사 간에] 문자 그대로의 동등함보다는 기능적 동등성에 초점을 맞춘다. 이러한 목표를 달성하기 위해 널리 쓰이는 번역 설계는 순방향 번역[16]과 역방향 번역이다. Brislin (1986)과 Hambleton 과 Patsula (1999)는 두 가지 번역 설계에 대한 정의, 강점 및 약점을 포함하여 전반적인 논의를 제공한다. 이 두 설계 각각은 결함이 있어 [어느 하나만으로는] 번역/번안 검사를 타당화하기에 충분한 증거를 제공하는 경우가 거의 없다는 것에 유의해야 한다. 역번역 설계의 주요 단점은 이 설계가 가장 좁은 형태로 구현될 경우

[15] 번안검사가 대상문화를 반영하는 정도가 높을 때 문화적 동질성이 확보된다.

[16] 순방향 번역은 줄여서 '순번역', 순방향이라는 맥락이 분명한 경우에 한해서 '번역'으로 서술하였다. 그렇지 않을 경우의 '번역'은 포괄적인 의미를 가진다.

검사의 대상언어 판 자체에 대한 검토가 수행되지 않는다는 것이다[17]. 그 결과로, 역번역이 최대한 쉽게 되도록 번역/번안이 이루어지는 경우가 자주 있고, 때로는 오히려 어색한 번역/번안 검사가 만들어지기도 한다.

두벌 번역[18] 및 조정 절차는 단일 번역의 특수성에 의존하는 단점과 위험을 해결하기 위한 것이다. 이 접근 방식에서 제3의 독립적 번역자 혹은 전문가 패널은 두벌의 순번역간의 불일치를 확인하고 해소하며 이를 단일판으로 조정한다. PISA와 같은, 대규모 문화간 평가 프로그램에서 두 가지의 다른 언어 판 (예, 영어 및 불어)을 각각 원본 검사로 하여 두벌의 번역본이 나오면 이를 조정하여 단일판이 얻어진다 (Grisay, 2003). 이 방법은 [두벌의 번역 간에] 불일치하는 부분을 파악하여 대상언어로 직접 검토하는 등 중요한 장점을 제공한다. 또한 두 개 이상의 언어로 된 원본들을 사용하면 원언어들의 문화적 특수성이 [번안 검사에] 미치는 영향을 최소화하는 데 도움이 된다.

언어 구조의 차이는 검사 번역에 문제를 가져올 수 있다. 예를 들어 Rotter 와 Rafferty (1950)가 영어로 개발한 잘 알려진 척도의 문항 중 하나인 "I like --- "; "I regret ---"; "I can't ---" 와 같은 문장완성형의 빈칸을 채워야 하는 검사가 있다. 그러나 이와 같은 문항 형식은 한국어에서는 부적절하며, 목적어는 주어와 동사 사이에 와야 한다[19]. 따라서 영어판처럼 불완전한 문장을 사용하면 한국 학생은 처음에 문장을 작성하기 전에 문장의 동사를 먼저 살펴야 하므로 응답 행동이 다를 수 있다.

이 문제에 대한 어떤 해결안에서도, 대상언어판은 형식의 측면에서 원본 검사와 다를 수 밖에 없다.

실무를 위한 제안. 검토자들의 판단 자료를 모아서 정리하는 것은, 지침이 준수되고 있는지 점검하는 데 특히 유용하다.

- Brislin (1986), Hambleton과 Zenisky (2010), Jeanrie와 Bertrand (1999)가 개발한 평정 척도를 사용한다. Hambleton과 Zenisky는 번안 과정에서 번역본에 대해서 점검해야 하는 25가지 측면들의 목록을 경험적으로 타당화해서 제공한다. Hambleton과 Zenisky(2010)의 목록 가운데에는 "번역 문항의 언어와 원본 검사의 언어 간에 난이도가 대등하고, 공유되는 성질이 있는가?" 그리고 "번역이 두 언어판 검사 문항의 난이도에 영향을 줄 정도로 문장상의

[17] 번역/번안된 것에 대한 검토없이, 그에 대한 역번역을 하여 원본검사와 비교하므로 번역/번안 검사 자체에 대한 검토는 생략된다.

[18] 두벌 번역은 번역자(팀)가 하나 아닌 둘인 경우로서 번역/번안 결과가 한벌이 아니고 두벌이다.

[19] 원문에서는 터키어의 예를 들었으나 번역과정에서 한국어에 대한 예로 바꾸었다.

변화(생략, 대체 또는 추가)가 있는가"가 있다.

- 현실적으로 가능하다면 복수의 번역 설계를 사용한다. 예를 들어, 전문가 패널이 두벌 번역 및 조정을 통해 작성한 다음, 재 점검을 위해 역번역 설계를 사용할 수 있다.
- 검사가 여러 문화권에서 사용되도록 고안된 경우, 향후에 문화권별로 번역/번안해야 하는 문제가 발생하지 않도록 처음부터 복수 언어판을 동시 개발하는 것을 고려한다. 검사의 동시 개발에 관한 자세한 정보는 Solano-Flores, Trumbull 과 Nelson-Barber (2002)에서 찾아볼 수 있다. 원본 검사를 설계할 때 적어도 향후 번역자에 의한 번역을 가능하게 하고 잠재적인 문제를 최대한 피할 수 있도록, 특별히 문화 특수적 사항이나 특이한 문항/반응양식 등을 피하도록 한다.
- 언어 간 구문 차이를 고려하면, 경직된 문장 구조의 경우 번역상의 문제 때문에 대규모의 국제 [교육] 평가에서 그리고 아마도 심리 검사에서도 그런 구조에 의존하는 형식은 피해야 한다.

검사개발-3 (6). 검사에서의 지시사항과 문항 내용이 모든 대상집단들에서 유사한 의미를 갖는다는 증거를 제공한다.

설명. 이 지침에서 요구하는 증거는 다양한 전략을 통해 수집할 수 있다 (예, van de Vijver & Tanzer, 1997). 이들 전략에는 (1) 현지 문화와 언어에서 원주민이라 할 수 있는 검토자를 사용, (2) 응답자들[20] 가운데 이중 언어 사용자들을 이용, (3) 검사를 평가하기 위해 현지에서의 조사 실시, (4) 수용가능성과 타당도를 향상시키기 위해 비표준화된 검사 실시를 해본다.

번안판을 소규모로 시행해보는 것은 좋은 방법이다. 소규모 시행은 검사 실시 및 자료 분석을 가능하게 하는 것은 물론, 가장 중요한 것은 실시자 및 응답자와의 면담을 통해 검사 자체에 대한 비평을 받을 수 있다. 상이한 언어배경의 내용 전문가, 또는 이중언어를 구사하는 내용 전문가를 활용한 번역 설계도 가능하다. 예를 들어, 이중언어를 사용하는 내용 전문가에게 두 가지 [원본, 번안판] 검사의 문항 형식과 내용에서 난이도가 유사한지 평가하도록 요구할 수 있다. 인지적 면접은[21] [용도가] 앞으로 기대되는 또다른 방법이다 (Levin et al., 2009).

---

[20] 번안검사를 실시하거나 그에 대한 의견을 구할 때의 응답자들을 의미한다.

[21] 인지적 면접은 인지작용이나 그 과정을 묻기 때문에 인지적 면접이라고 한다. 번안검사제작시 인지적 면접은, 번안 검사의 본검사가 제작되기 전에 일종의 예비검사 단계에서 검사중 또는 검사후 8~12명 정도의 응답자들에게 개방형 질문이나 언어적 탐침(probing) 질문을 통해 검사를

실무를 위한 제안. 이 지침을 구현하기 위한 몇 가지 제안들이 있다.

- 검사의 번역/번안을 평가하기 위해 현지 문화와 언어에서 원주민이라고 할 수 있는 검토자를 이용한다.
- 검사에서의 지시사항과 문항측면에서 두가지 언어판 간 동등성에 대한 의견을 구하고자 이중언어 응답자 표본을 사용한다.
- 현지에서의 설문조사를 통해 검사를 평가한다. 이러한 소규모 시행은 매우 가치가 있다. 검사 문항에 대한 응답자의 실제 응답보다 실시자 및 응답자의 의견이 더 가치 있는 경우가 많으므로 검사 실시 후 반드시 이들을 면담한다.
- 수용가능성과 타당도를 높이기 위해, 번안된 검사를 몇번 실시해 본다[22] [매번 지시사항을 달리해서 실시해야 한다]. 각 실시에서 유사한 지시사항을 사용하면 대상언어/문화 집단이 어디에서 잘못된 이해를 하였는지 알 수가 없다.

검사개발-4(7). 문항 형식, 평정 척도, 점수부여 범주, 검사상의 관례[23], 실시 방법 및 기타 절차가 의도된 모든 대상집단에 적절하다는 증거를 제공한다.

설명. 5점 척도나, [컴퓨터 사용검사에서 마우스로] "끌어다 놓기" 또는 "올바른 것에 모두 대답하기", 심지어 "단 한 개의 대답만 선택하기"와 같은 문항의 형식은 이전에 이러한 형식을 사용해보지 못한 응답자들에게 혼란을 줄 수 있다. 심지어 문항의 배열, 그래픽의 사용이나 급격하게 확산되는 컴퓨터화 된 문항의 형식도 참가자들에게 혼란을 줄 수 있다. 이러한 유형의 오류에 대한 많은 예가 미국에서 아동에 대한 표준화된 검사를 컴퓨터로 옮길 때 발생하였다. 이러한 문제는 연습 문제를 통해 대부분의 아이들이 극복할 수 있다. 응답자들은 이러한 새로운 문항 형식에 친숙해야 하는데, 그렇지 못하면 개인 및 집단 검사 결과를 왜곡할 수 있는 검사 편향의 원인이 된다.

평가하기 위한 질적 접근이다. 응답과정에서 어떤 인지 작용이 있는지 파악하고자 "그 설문에서 묻는 것은 무엇이었나요?" 또는 "왜 당신의 건강이 '보통'이라고 했는지 이야기 해주시겠습니까?"등의 질문을 한다.

[22] 물론 이것은 소규모의 비표준화된 실시를 말한다.

[23] 검사의 문항, 채점, 점수체계, 실시의 구체적 사항등 검사전반에서 일반적으로 이루어지는 부분을 가리킨다.

컴퓨터 기반 검사와 관련하여 최근에 발생하는 문제가 있다. 컴퓨터 기반 검사 플랫폼[24]에 익숙하지 않은 경우, 응답자가 그에 익숙해지도록 개별지도 또는 지침서가 있어야 의미있는 점수를 제공할 수 있다.

실무를 위한 제안. 이 지침이 준수되는지 평가하는 데는 질적 증거와 양적 증거가 각기 역할을 한다. 번안된 검사에서 점검해야 할 몇 가지 사항은 다음과 같다.

- 연습 문제가, 응답자들이 올바른 응답을 할 수 있는 수준까지 또는 검사 자료에 대한 그들의 숙달도[25]를 반영하는 정도까지 끌어 올리기에 충분한지 점검한다.
- 응답자가 검사 진행과정에 들어 있는 새로운 문항 형식이나 실시방식(예, 컴퓨터기반 실시)을 숙지하도록 한다.
- 검사상의 관례(예, 검사내 도표의 배치 또는 답안지에 답안 표시)가 응답자에게 잘 이해되는지 확인한다.
- Jeanrie와 Bertrand(1999) 그리고 Hambleton과 Zenisky(2010)가 개발한 평가 척도가 도움이 된다. 예를 들어 Hambleton과 Zenisky는 "두가지 언어판 간에 물리적 배열을 포함한 문항 형식이 [원문과 번안판 간에] 동일한가?" 그리고 "단어 또는 어구의 강조(진하게, 기울임, 밑줄 등)가 원본 문항에서 사용된 경우, 번역된 문항에서도 그렇게 강조되었는가" 등의 질문을 통해 검토한다.

검사개발-5 (8). 문항분석, 신뢰도 평가 및 소규모의 타당도 연구를 통해 번안검사에 수정할 수 있도록 예비검사를 실시하여 자료를 수집한다.

설명. 많은 시간과 비용이 소요되는 대규모 검사[26]를 통해 점수의 신뢰도 및 타당도 연구 및/또는 규준작성[27] 연구를 시작하기 전에 번안된 검사의 심리측정학적 증거를 확인하는 것이

---

[24] 컴퓨터 기반 검사 플랫폼은 컴퓨터의 도입으로 가능해진 검사관리 시스템을 가리킨다. 즉, 검사 관리자는 문항은행에서 문항을 선별하여 검사를 구성하고 응시자의 점수를 관리할 수 있고, 응시자는 개인 번호를 이용하여 컴퓨터 화면상에서 검사를 실시하고 즉시 점수를 알수있는 시스템이다.

[25] 응답자의 역량수준을 가리킨다.

[26] 이런 용도의 대규모 검사를 본검사라고도 한다

[27] 규준참조검사(상대평가용)의 경우 규준(norm)이 작성되지만, 준거참조검사(절대평가용)의 경우

중요하다. 검사 점수의 신뢰도와 타당도에 대한 초기 증거를 제공하기 위해 수행될 수 있는 많은 심리측정학적 분석이 있다. 예를 들어, 검사 개발 초기 단계에서, 최소한의 적당한 표본 크기(예: 100)와 문항 정보는 특정 검사 문항의 기능파악에 매우 필요한 자료를 제공할 수 있다. 다른 문항에 비해 매우 쉽거나 어렵거나 또는 낮은 변별력을 가진 문항들의 결함에 대해 검토할 수 있다. 다지선다형 문항에서는 오답선지들의 효과성을 조사하는 것이 적절할 것이다. 문항분석을 통해 문항의 문제들이 발견되고 수정될 수 있다. 또한, 문항 분석을 위해 수집된 자료에서, 알파 계수 또는 오메가 계수 (McDonald, 1999)[28] 가 산출되고 검사 개발자가 원본 및 대상언어판 검사의 적절한 길이를 결정하는데 귀중한 정보를 제공한다.

경우에 따라, 번안의 특정 측면에 대한 문제가 여전히 있을 수 있다. 검사에서의 지시사항이 온전히 이해될 수 있는가? 새로운 언어와 문화에서의 응답자들을 효과적으로 안내하려면 지시사항이 달라져야 하는가? 컴퓨터 기반 검사는 번안검사의 대상집단에서 특정한 소집단(예: 낮은 사회경제적 지위의 응답자)에게 문제를 발생시킬 것인가? 사용 가능한 시간에 너무 많은 문항들이 제시되고 있지 않는가? 이 모든 문제들과 더 많은 것들에 대하여, 어느정도 규모(modest-sized)[29] 있는 타당화 연구를 통해 답할 수 있다. 이러한 연구의 목적은 다음 단계로 진행할 수 있는지에 대한 결정을 내리는데 충분한 자료를 수집하는 것이다. 만약 다음 단계로의 진행이 결정되면, 일련의 상당히 큰 연구들이 계획되고 수행된다(예, 차별기능문항 연구, 검사의 요인 구조 연구)[30].

실무를 위한 제안. 여러 가지의 기본적인 분석들을 수행할 수 있다.

기준점수(cut-score)가 작성된다.

[28] 알파 계수는 흔히 Cronbach 알파 또는 정확하게는 Cronbach-Guttman α 계수라고 하고 고전검사이론에서 정의된 신뢰도의 추정치이다. 그에 반해 오메가(ω) 계수는 McDonald가 제시한 신뢰도 추정치로서 요인분석의 결과에 고전검사이론의 정의를 적용해서, 요인점수의 신뢰도를 구한 것이다. 이것은 측정의 오차가 포함된 검사점수 자체의 신뢰도가 아니고, 그 오차가 배제된 요인점수의 신뢰도이므로 좀 더 엄격한 개념이다.

[29] 적어도 100명보다는 크지만, DIF 연구나 요인구조 연구에 필요한 표본크기보다는 작은 규모를 의미한다.

[30] 이 지침서에서는 '상당히 큰 연구들'은 본검사에서의 연구를 의미한다. 그러나, 검사 제작의 실제에서 이 부분 전체를 본검사에서 진행하기에는 너무 늦다. 예비검사 단계에서 어느정도 실시되어야, 본 검사에서 확인 및 정교화 할 수 있다.

- 고전적인 문항분석을 수행하여 문항 평균점수 및 문항 변별도[31]를 구한다. 그리고 다지선다형 문항이나 유사한 선택형 문항에 대해서는 오답선지분석[32] 을 수행한다.
- 신뢰도 분석을 실시한다(예, 2점척도로 채점되는 문항에서는 KR-20[33], 또는 다분척도로 채점되는 문항에서는 알파 또는 오메가 계수).
- 필요에 따라 한 두가지 연구를 수행하여 번안 검사의 타당도에 대한 통찰을 얻는다. 예를 들어, 번안 검사가 컴퓨터를 통해 실시된다고 가정하면, 검사 실시의 방식(즉, 지필과 컴퓨터로 실시)을 평가하는 연구를 진행해야 한다. 검사 지시사항에서 응답자들이 모든 문항에 답하도록 요구할 경우, 이를 달성하기 위한 최선의 지시사항이 무엇인지 파악하기 위한 연구가 필요할 수 있다. 응답자들에게 필요한 정보가 없을 경우 추측하여 답을 하라고 권해도, 응답자에 따라서는 모든 문항에 답하는 것이 놀라울 정도로 어려운 것임을 연구자들이 발견하고 있다.

3. 검수 지침

검수 지침은 본격적인 타당도 연구[34]에서의 경험적 분석에 대한 것이다.

검수-1 (9). 검사의 대상집단으로서의 특징을 가지며 경험적 분석을 위한 충분한 크기의 표본을 선정한다.

설명. 자료수집 설계는 상이한 언어판 간의 규준 작성 (필요한 경우)및 [검사] 동등성을 확인하고, 타당도 및 신뢰도 연구, 차별기능문항 연구를 실시하기 위해 자료를 수집하는 방식을 말한다. 자료의 첫 번째 요건은 안정적인 통계 정보가 가능할 정도로 충분히 큰 표본이어야 한다는 것이다. 이는 어떠한 연구에도 적용되지만, 번안 검사의 타당화 연구에 특별히 중요하다. 왜냐하면, 검사의 동등성 및 문항 동등성을 확립하는데 필요한 통계적 기법(예, 확인적 요인 분석, 편향될 수 있는 문항의 식별에 대한 IRT 방법)은 모형의 모수를 안정적으로 추정할 수 있을

---

[31] 문항변별도는 문항점수와 검사점수(문항 전체를 통괄하는 점수)간 관계의 강도이다.

[32] 오답선지 분석은 오답선지들에 대한 반응분포가 무선(random)적이지 않고 어느 한쪽으로 쏠리는지를 분석하는 것이다.

[33] KR-20는 이분 문항들에 알파계수 공식을 적용한 것과 동일한 결과를 준다.

[34] 이러한 연구는 예비검사에서 대체로 파악된 내용들을 다시 한번 확인 및 정교화 하는 절차가 되므로 본검사 자료에 기반하여 이루어 진다.

만큼의 큰 표본이 있을 때 적용될 수 있기 때문이다. (권장되는 표본의 크기는 모형의 복잡성과 자료의 질적수준에 따라 달라진다).

또한, 본격적인 타당화 연구를 위한 표본은 검사 대상의 모집단을 대표해야 한다. 적절한 통계적 설계 및 분석을 위한 안내서로써 van de Vijver와 Tanzer (1997)의 주요 논문과 Byrne (2008), Hambleton 등 (2005), van de Vijver와 Leung (1997), 그리고 Byrne과 van de Vijver (2010)의 방법론적 공헌에 주목할 수 있다. Sireci (1997)는 여러 언어판 검사들을 공통의 척도로 연결하는 데 있어 문제점과 쟁점에 대한 논의를 제공했다.

실제에서 때로는, 번안 검사의 대상언어 집단은 원본검사의 언어집단에 비해 훨씬 더 낮은 혹은 더 높은 점수를 내기도 하고, 집단 내 동질성이 원 언어 집단보다 높거나 낮은 경우가 있다[35]. 이는 신뢰도 및 타당도 연구 등 특정 분석 방법에 대해 큰 문제를 일으킨다. 이를 위한 하나의 해결책으로 대상언어 집단의 표본과 일치하도록 원본검사의 집단에서 하위집단을 선택하는 것이다.[36] 이렇게 대응[37] 되는 두 표본을 비교하면 앞서 관찰된 분포모양(shape)의 차이[38]로 인해 검사 결과에서 발생하는 차이를 모두 제거할 수 있다(Sireci & Wells, 2010 참조). 예를 들어, 검사 구조의 비교는 일반적으로 공분산을 포함하는데, 공분산은 집단 내 점수의 분포에 따라 변한다. [원본검사 집단과 번안검사 집단 간에] 대응되는 표본을 사용하면, 검사 결과에 작용하는 [오염변수로 인한] 점수의 분포가 두 표본 간에 대응하므로, 표본 간 결과의 차이에서 [오염변수로 인한] 점수 분포의 차이를 배제할 수 있다.

원본언어집단과 대상언어 집단 간에 점수 분포가 달라서 생기는 문제를 설명하는 또 하나의 예가 있다. 검사 점수의 신뢰도가 원본언어 집단에서 .80인 반면, 대상언어 집단에서 .60이라고 가정해보자. 이 차이는 큰 문제로 여겨질 수 있고, 번안검사의 적절성에 대한 의문이 제기될 수 있다. 그러나 신뢰도는 검사와 대상집단이 결합된 특성임이 자주 간과되고 있다(McDonald, 1999). 신뢰도는 진 점수의 분산(대상집단 특성)과 오차분산(검사 특성)에 따라 달라질 수 있기 때문이다. 따라서, 원본언어 집단에서 진점수 분산이 커지면 오차분산에 변화가

---

[35] 이런 내용은 집단간 분포모양(shape)의 차이를 의미한다.

[36] 이것은 원본검사 집단에 비해 번안검사 집단이 더 동질적(분산이 훨씬 작음)일 경우를 가정하는 것이다.

[37] 여기서 "대응"은 앞 문장에서의 "일치"와 같은 의미로 쓴 것인데, 이 문단의 첫 문장에서 나온 집단 간 점수 분포의 차이를 가져 올 수 있는 오염변수의 관점에서 동일한 수준이 될 수 있도록 하다는 의미이다.

[38] 집단 간에 오염 변수들이 동등한 수준으로 대응되지 않아서 발생한 점수 분포의 차이를 의미한다.

없어도[39] 신뢰도는 더 큰 값이 될 수 있다. 그래서 McDonald (1999)는 표본 간의 비교에서 신뢰도 아닌 측정의 표준오차(오차분산의 제곱근)[40]가 더 적합한 수치라는 것을 보여준다. [이러한 상황에서] 신뢰도 계수를 사용하겠다면, 원본언어 집단에서 [대상언어 집단에] 대응하는 표본을 추출하여 검사의 신뢰도 계수를 다시 계산하는 것이다[41].

다집단 간 확인적 요인 분석(CFA)으로 측정의 동일성을 검증하는 현대적인 방법은 잠재속성[42]의 분포가 서로 다른 표본 간에 평가를 가능하게 한다. 이러한 모형에서, 문항의 요인계수(factor loading) 및 절편(intercept)과 같은 측정 모형의 모수는 집단 간 동일하다고 가정하는 가운데, 잠재속성의 평균(mean), 분산(variance) 및 공분산(covariance)은 집단마다 다른 값으로 추정될 수 있다[43]. 이를 위해 집단별로 대규모 표본의 사용을 고려하게 되고 측정된 구성개념들이 집단 간에 서로 다른 분포를 가질 수 있다는, 보다 현실적인 분석 계획을 가능하게 한다.

실무를 위한 제안. 거의 모든 연구에서, 표본[44]을 설명할 때 두 가지 제안이 있다.

- 하나의 언어판에 대한 검사에서 편향 가능성 있는 문항을 식별하기 위한 연구에 최소한 200명의 표본이 필요하다는 점을 고려하여(Mazor, Clauser, & Hambleton, 1992; Subok, 2017), [집단별로] 가능한 한 큰 표본을 수집한다. 문항반응이론을 이용한 분석과 모형 합치도의 연구를 실시하기 위해서는 최소 500명의 응답자 표본이 필요하다(Hambleton, Swaminathan, & Rogers, 1991; Hulin, Lissak, & Drasgow, 1982). 반면에 검사의 요인 구조를

[39] 오차분산은 동일해도 진점수 분산이 커지는 경우는, 검사의 제반 관행이 유지되는 가운데 (오차분산 불변), 구성개념을 충분히 폭 넓게 측정하는(진점수 분산이 커짐) 검사개발이 이루어지는 경우이다. 신뢰도가 높아진다.

[40] 검사 점수 분산이 100이고 신뢰도가 .75이면 100(1-0.75)=0.25가 오차분산이고 측정의 표준오차는 0.25의 제곱근인 0.5가 된다.

[41] 물론 원본 검사집단에 비해 번안 검사집단이 더 동질적일 때 가능한 대안이다.

[42] 요인, 이론적 변수라고도 하며, 검사에서 측정의 대상이 되는 구성개념에 대한 경험적 분석시에 사용되는 용어이다.

[43] 측정모형의 동일성(측정틀 동일성, 요인계수 동일성, 측정오차 분산 동일성, 절편 동일성)이 성립하고 이론변수의 분산/공분산들이 집단간에 동일하면 요인점수 수준에서 집단간 비교를 하는 요인(잠재속성) 평균비교가 가능하다. 물론 집단 간 이론 변수(요인)의 분산/공분산들의 동일성이 위반되면 그 수준에서 집단별 해석을 하고 요인평균 비교까지는 가지 않는다.

[44] 여기서 표본은 언어판 별 표본이다.

조사하기 위한 연구는 약 300명 혹은 더 많은 수의 표본 크기를 필요로 한다(Wolf, Harrington, Clark, & Miller, 2013)[45]. 물론 더 작은 표본으로 분석할 수도 있지만[46], 분명한 제1원칙은 가능한 한 큰 표본을 구하는 것이다.

- 가능한 경우 응답자 집단에서 대표성 있는 표본을 선정한다. 대표성 없는 표본의 결과를 가지고 [모집단에] 일반화하는 데는 제한이 있다. 방법론적 이유(예, 집단 간에 점수분포가 다름)로 발생하는 집단 차이를 제거하려면, 대상언어 집단에 대응하는 표본을 원본언어 집단에서도 구하는 것이 좋은 방법인 경우가 자주 있다. [집단 간에] 측정의 표준오차를 비교하는 것이 신뢰도 계수를 비교하는 것보다 더 적절할 수 있다.

검수-2 (10). 집단 간에 구성개념의 동등성, [측정] 방법 동등성 및 문항 동등성에 대한 관련 통계 수치를 제공해야 한다.[47]

설명. 검사의 원본언어 집단과 대상언어 집단 간 구성개념의 동등성을 확립하는 것이 중요하지만, 이것이 중요한 경험 분석의 전부는 아니다. 구성개념의 동등성(선행조건-2)과 측정방법의 동등성 (선행조건-3)은 지침서의 앞부분에서 간략하게 설명되었다[48]. 연구자들은 또한 문항의 동등성 수준을 제시할 필요가 있다. 문항의 동등성은 "차별기능문항(DIF)분석"이라는 이름으로 연구된다. 일반적으로 DIF는 서로 다른 (언어/문화) 집단에 속한 두 명의 응답자가, 측정된 속성(요인, 구성개념)의 수준에서는 동일하지만 검사 문항에 대한 반응 확률[49]이 서로 다른

---

[45] 이들의 연구에서는 확인적 성격의 요인분석이고, 요인수효가 3개였다. 요인수효가 많고 추정해야하는 모수(parameter)가 증가하면 당연히 더 큰 표본이 필요하다. 탐색적인 요인분석에서는 확인적 요인분석보다 더 큰 표본이 필요하고 사용되는 측정치(문항점수 또는 소검사 점수)수효의 5~10배의 표본이 필요하다.

[46] 요인분석시에 자료내 이론적 구조가 "분명할수록" (확인적 성격의 요인분석) 요구되는 표본크기는 작아질 수 있는데, 경우에 따라 측정치 수효의 5배 미만도 가능하다.

[47] 집단간에 이러한 동등성 검토는 확인적 요인분석의 이론모형과 측정모형이 집단간에 동일한지의 검정을 통해서 가능하다.

[48] 선행조건-2와 선행조건-3은 모두가 질적평가 즉, 전문가 판단에 의존한다. 그러나 검수-2(10)은 그러한 판단을 뒷받침하는 통계수치를 요구한다.

[49] 이 반응 확률에 따라 DIF 측정치가 산출 된다. 검사에서 측정하는 구성 개념의 수준에서는 두 집단이 동일한데 문항점수가 다르게 나오면 그 문항은 집단에 따라 기능이 달라지는 문항 즉, DIF

경우에 존재한다. 검사점수가 집단 간에 차이를 보일 수 있지만 이것만으로는 문제라고 할 수가 없다[50]. 반면에 구성개념 수준(일반적으로, 검사 총점 또는 총점 계산시, DIF 가능성이 검토되는 해당문항을 제외한 수정된 검사총점)에서는 동일하지만 문항수준에서 집단 간의 차이가 있을 경우 해당 문항은 DIF를 나타낸 것이다. [문항동등성 또는 반대 개념으로서의 차별기능문항에 대한] 이러한 분석은 검사의 각 개별 문항에 대해 수행되어야 한다. DIF가 드러난 이후에 그 이유를 이해하려는 노력이 뒤따르고, 그 판단적 검토[51]를 통해 문항의 일부에서 결함이 판명되면 수정되거나 검사에서 제거될 수 있다.

DIF의 잠재적 원인으로 검토되어야 할 두 가지는 번역상의 문제와 문화적 차이다. DIF는 구체적으로, (1) 원본언어에서 대상언어로 번역 시 사용된 어휘에 대한 친숙성, 문항 난이도의 변화 및 의미가 동등하지 않게 되는 것 등과 같은 번역상의 차이 그리고 (2) 문화적 맥락의 차이에 기인할 수 있다(Allalouf, Hambleton, & Sireci, 1999; Ercikan, 1998, 2002; Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; Ercikan, Simon, & Oliveri, 2013; Li, Cohen, & Ibarra, 2004; Park, Pearson, & Reckase, 2005; Scheuneman & Grima, 1997; Sireci & Berberoğlu, 2000; van de Vijver & Tanzer, 1997).

번역 과정에서, 대상언어에서는 잘 사용되지 않는 어휘가 사용될 가능성이 있다. 번역본에서의 의미는 같겠지만, 그 단어는 어느 한 문화보다 다른 문화에서 더 보편적일 수 있다. 또한 문장의 길이, 문장의 복잡성, 쉬운 어휘나 어려운 어휘의 사용으로 인해 번역된 문항의 난이도가 변할 수 있다. 또한 문장의 일부분을 삭제하거나, 부정확한 번역, 해당 어휘가 대상언어에서 둘 이상의 의미를 가지는 경우, 일부 단어들의 의미가 문화 간에 동등하지 않은 느낌을 주는 등 대상언어에서의 의미가 변할 수 있다. 무엇보다, 상이한 언어에 따른 문화 차이는 문항의 차별적 기능을 가져올 수 있다. 예를 들어, "햄버거"나 "금전 등록기"와 같은 단어는 문화가 다르면 이해되지 않거나 다른 의미를 가질 수 있다.

문항들이 언어 및/혹은 문화 집단에 따라 그 기능을 달리하는지를 확인하기 위한 분석 절차가 최소한 4개 있다. (a) IRT-기반 절차 (예, Ellis, 1989; Ellis & Kimmel, 1992; Thissen, Steinberg, & Wainer, 1988; 1993 참조), (b) Mantel-Haenszel (MH) 절차 및 수정된 MH 절차(예, Dorans & Holland, 1993; Hambleton, Clauser, Mazor, & Jones, 1993; Holland & Wainer, 1993; Sireci & Allalouf,

---

문항이다.

[50] 집단간에 검사점수가 다를 때, 구성개념 수준에서의 집단간 차이인지, 각 문항수준에서 DIF의 결과로 나온 것인지 알 수가 없다.

[51] 판단적 검토란 경험자료를 수집해서 분석하는 것이 아니라 DIF 분석결과를 보고 DIF를 보이는 이유에 대하여 질적인 추론을 하는 것을 의미한다.

2003 참조), (c) 로지스틱 회귀 (LR) 절차 (예, Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990), (d) 제약 요인 분석 (RFA: restricted factor analysis) 절차가 있다(Oort & Berberoğlu, 1992)[52].

IRT-기반 방법에서는, 두 언어집단에서의 응답자들을 잠재속성 점수에 근거하여 대응시킨다[53]. MH 그리고 LR 방법론에서는, 집단 간에 응답자들의 문항 수행[문항점수]을 비교하기 전에 응답자들을 대응시키는 기준으로서 관찰된 검사점수 또는 다른 여러 추정치를 사용한다. 이런 절차들에서 가장 널리 쓰이는 대응기준은 문항점수들을 합산한 검사점수이지만[54], 다른 방식의 추정치 예를 들어, 요인 분석에서 얻은 요인 점수가 사용될 수도 있다. 이러한 문항분석을 통해 의심스러운 문항들을 삭제함으로써 검사에서 점차 오염이 제거된다. 대응기준은 DIF를 적절하게 평가할 수 있을 정도로 타당하고 신뢰로운 것이어야 한다. RFA에서, 각 문항은 잠재 변수(구성개념을 나타내는 요인) 뿐만 아니라 집단 소속 변수(DIF를 유발할 것으로 보이는 변수)를 독립변수로 하여 회귀 분석[55] 을 실시한다. 각 문항과 구성개념 요인 간의 계수는 자유모수로 추정되며, 각 문항과 집단 소속 요인 간의 계수는 모두 0으로 고정(DIF없는 모형)된 영가설 모형의 합치도를 평가한다. 영가설 모형의 합치도가 좋으면 DIF가 없는 것이다. 합치도가 낮으면 문항과 집단 소속 요인 간 계수 가운데 모형의 합치도를 가장 높게 향상시킬 것 같은 고정모수(0으로 고정)를 자유모수로 추정하는 수정모형을 합치시켜본다. 이 수정 모형의 합치도가 유의하게 향상되면 그 문항은 DIF 문항으로 판정한다.

검사에서 측정되는 요인이 복수일 때, 적절한 대응 기준을 찾는 것이 어려워진다(Clauser, Nungester, Mazor, & Ripley, 1996). 요인분석의 결과로 복수의 요인들에 대해 요인점수들이 산출되는데, 이러한 다변량적인 대응기준을 사용하면 문항수준에서 DIF 해석이 달라질 수가 있다. 따라서 이 지침에서는, 검사가 다차원[다요인]일 경우 연구자가 다양한 대응

---

[52] 제약 요인분석에서는 측정되는 구성개념을 요인으로 하는 확인적 요인분석 모형에 집단 소속변수(예: 원본언어집단 소속-0, 대상언어집단소속-1)를 하나의 요인으로 추가한 다음, 그 요인과 각 문항간 요인계수를 일단 모두 0으로 제약한다. 다음은, 그 제약 모수들을 하나씩 자유모수로 해줌에 따라 모형의 합치도나 추정치가 변하는 정도를 보고서 각 문항의 DIF 정도를 파악한다.

[53] 집단간에 잠재 점수가 같은 사람끼리 대응시킨다는 것인데, 구성개념 관점에서 동등한 수준이 이루어지는 것이다. 이러한 대응 작업에서 사용되는 매개변수를 대응기준이라고 한다.

[54] 검토되는 검사가 시험이라면 문항점수의 합산이 검사점수가 되고 평정척도(예: 5점 척도, 7점척도 등)라면 합산내지 평균이 검사점수가 된다.

[55] 요인들(구성개념, 집단소속)과 문항점수간 회귀분석이 곧 요인분석이다(여기선 확인적 요인분석).

기준을 사용하여 DIF 문항들을 찾아내고, 다수의 대응 기준에 걸쳐서 일관되게 DIF 가능성을 보이는 문항들을 찾아볼 것을 제안한다. [대응기준 중 하나로서] 다변량적 대응기준은 언어/문화 차이가 있는 집단 간에 가능한 DIF 문항의 수를 적게 나타나게 할 수가 있다[56].

문항의 동등성을 평가하는 위의 방법론들에서 요구되는 표본의 크기가 서로 다를 수 있다. MH, LR 및 RFA는, 모수 추정을 타당하게 하기 위해, 더 큰 표본이 있어야 하는 IRT-기반 기법보다 상대적으로 작은 표본으로 신뢰롭고 타당하게 평가할 수 있는 모형이다. [표본크기에 더해서] 문항 반응 자료의 유형 또한 고려해야 한다. MH, LR, 그리고 RFA 방법은 이분 채점되는 자료에 사용할 수 있다. 다분 반응 [채점]자료에는 일반화된 MH 절차와 같은 다른 방법들이 가능하다[57].

이 지침에 따라 검사를 평가할 때, 연구자는 번안 검사에서 방법 편향의 원천이 되는 부분이 있는지 찾아야 한다. 방법 편향의 원천으로는 (1)검사참가자의 동기부여 수준차이, (2) 심리검사에 대한 경험의 차이, (3), 언어 집단에 따른 검사의 신속성 차이[58], (4) 언어 집단 간 검사 반응 형식에 대한 친숙성의 차이, (5) 반응 방식의 차이 등이 있다. 응답의 편향은 PISA 결과를 해석하는데 주요 관심사였고 몇몇 연구에서 주목을 받았다.

마지막으로, 또 하나 중요한 내용은, 이 지침에서는 연구자들이 구성개념의 동등성을 평가할 것을 요구한다는 것이다. 검사의 원본언어판과 대상언어판 간에 걸쳐 구성개념의 동등성을 평가하기 위한 통계적 방법에는 탐색적 요인 분석(EFA), 확인적 요인 분석(CFA), 다차원 척도법 (MDS), 논리적 관계망 비교 등 적어도 4가지 이상이 있다 (Sireci, Patsula, & Hambleton, 2005).

van de Vijver와 Poortinga (1991)에 따르면, 요인 분석 (EFA 와 CFA)은 한 문화에서의 구성개념이 다른 문화에서 동일한 형태와 빈도로 발견되는지 여부를 평가하기 위해 가장 자주 사용되는 통계 기법이다. 1991년에 제시된 이 명제는 통계적 모형들이 상당히 진전된 오늘날에도 여전히 사실이다(예, Byrne, 2008, Hambleton & Lee, 2013 참조). EFA 방법으로는 서로 다른 요인 구조를 [통계적 검정을 통해] 비교하는 것이 어렵고, 구조의 동등성을 결정하는데 일반적으로 합의된 규칙이 없기 때문에, CFA(예, Byrne, 2001, 2003, 2006, 2008 참조), 가중 다차원

---

[56] 다변량적 대응 기준 만을 사용할 때 주의할 점이다.

[57] RFA는 확인적 요인분석을 용용한 방법으로 추정방법을 정확하게 선택하면 측정치가 이분채점이거나 다분채점되거나에 무관하게 사용될 수 있다. RFA에서는 측정오차를 통제하면서 통계적 검정이 이루어지므로 MH, LR, 일반화 MH방법에서 측정오차를 통제하지 못하는 어려움을 극복하게 된다.

[58] 동일한 검사지만 검사 실시가 집단에 따라 빠르게/느리게 진행되는 것을 의미한다.

척도법(WMDS)[59]과 같이 동시에 여러 집단을 한 모형에서 평가할 수 있는 통계적 방법이 권장된다(Sires, Harter, Yang, & Bhola, 2003).

기존 연구에서, 원본 검사의 요인 구조가 번안검사에서도 일관성이 있는지에 대하여 CFA 방법을 사용한 연구가 많이 이루어 졌다(예, Byrne & van de Vijver, 2014). CFA는 여러 집단을 동시에 모형화할 수 있고 모형의 합치도에 대한 통계적 검정은 물론 판단적 지수가 제공되기 때문에[60], 번안 검사들간 구조 동등성을 평가하는데 매력적인 방법이다(Sireci 등, 2005). 여러 집단을 함께 모형화할 수 있는 능력은 특별히 중요한데 그 이유는, 검사를 여러 언어로 번안하는 것이 보편화되고 있기 때문이다(예, 지능 검사중 어떤 것은 현재 100개 이상의 언어로 번역/번안되고 있으며, TIMSS와 OECD/PISA의 경우 검사가 30개 이상의 언어로 번안되고 있다). 그러나 CFA에서, 하나의 측정변수는 1개의 요인에 대해서만 지표(indicator)가 되도록 요구할 경우 복잡한 다차원[다요인구조] 척도의 자료에 대한 합치도가 낮아지는 경우가 많다. 그에 따라 성격 자료 또는 더 복잡하고 상호 관련된 변수들에 대해서는 탐색적 구조방정식 모형(ESEM)이 점점 더 인기를 끌고 있다(Asparouhov & Muthén, 2009)[61].

WMDS는 서로 다른 언어판의 검사들간 구성개념 동등성을 평가하기 위한 또 다른 매력적인 방법중의 하나이다. EFA와 마찬가지로 WMDS 분석에서는 검사의 구조를 사전에 지정할 필요가 없지만, CFA와 마찬가지로 복수의 집단을 분석할 수 있다(예, Sireci et al., 2003).

Van de Vijver 와 Tanzer (1997)는 다문화 연구자들이 다루는 검사에서 각 문화에 따른 번안판의 신뢰도를 검토하고 각 문화 집단에서의 수렴타당도와 변별타당도가 있는지를 조사해야 한다고 제안했다[62]. 이러한 연구는 상당한 표본 크기를 요구하는 검사 구조의 연구보다 더 실용적일 수 있다.

---

[59] WMDS(Weighted Multi-Dimensional Scaling)는 MDS 분석을 할 때 모수에 가중치를 부여함으로써 집단간 구조를 비교가능하게 한다.

[60] 합치도 가운데 χ^2가 검증적 합치도라면 CFI, TLI, RMSEA, 그리고 SRMR등은 판단적(비검증적)합치도이다. '검증적'이란 통계적 검정이 되는 것을 의미하고, '판단적'은 통계적 검정 없이 합치도의 값을 보고 모형이 자료에 합치하는 정도를 판단한다는 의미이다.

[61] ESEM(exploratory structural equation modeling)은 원래가 확인적 정신이 중심이 되는 구조방정식 모형에 탐색적 정신을 가미한 것인데, 국내에서도 응용 논문들이 종종 출판되고 있다(예, 한국산업 및 조직심리학회지, 한국임상심리학회지).

[62] 수렴타당도와 변별타당도를 검토하는 것은 하나의 집단에서 해당 구성개념이 타당한지를 파악(구성개념 타당화)하는 것이 핵심이다.

그러나 하나의 검사에 대한 두 언어판 간에 응답자의 검사수행을 비교하는 것이 검사의 번역/번안에 반드시 수반되는 목표는 아니라는 것을 알아야한다. 아마도 [검사점수 자체 보다는] 구성개념 측면에서 언어 집단 간 차이를 평가하는 것이 목표일수가 있다. 이 경우, 제2언어 [대상언어] 집단에서 [번안] 검사의 타당도에 대한 면밀한 검토는 필수적이지만[63], 두 언어판 검사 양식(form)이 동등한 지 아닌지는 그다지 중요하지 않을 수 있다. 따라서, 이 지침이 중요한지 아닌지는 대상언어집단에서의 검사 목표에 따라 달라질 수 있다. 예로서, PISA 또는 TIMSS에서 사용되는 검사들은 그 결과를 여러 나라 학생들의 성취도와 비교하고자 하는 것이기 때문에 내용이 고도로 유사하다는 증거가 필요하다. 그러나 한국 연구자들이 우울증을 연구하거나 한국 상담자들이 한국인 내담자의 우울정도를 평가하기 위해 영어에서 한국어로 번역된 우울 진단도구를 사용할 때 문항의 내용이 [영어권과 한국 간에] 고도로 유사할 것을 요구하지는 않을 것이다[64]. 그보다는 한국에서 사용될 우울 척도에 대한 타당화가 더 필요할 것이다.

이 지침은 검사의 번안이 완료된 이후에, 통계적 방법을 사용해서 검토될 수도 있다. 예를 들어, 상이한 문화집단들 간에 [검사에서 측정되는] 구성개념과 무관한 다른 중요 변수의 측면에서 차이가 있다면, 포괄적인 설계와 통계 분석을 사용하여 이러한 '오염' 요소를 통제할 수 있다. 공분산 분석, 블록내 무선화 설계[65] 및 기타 통계 기법(회귀 분석, 부분 상관계수 등)을 사용하여 집단 간 원치 않는 차이가 가져오는 효과를 통제[66]할 수 있다.

---

[63] 우선은, 하나의 대상 언어 집단에서 타당도 있는 번안검사라야 다른 언어집단과의 비교가 의미 있기 때문이다.

[64] 이 부분은 영어와 중국어 번역본에 대한 원래 내용을 영어와 한국어에 대한 내용으로 대체하였다. 임상/상담 장면 그리고 주제에 따라서는 산업/조직 장면에서 번역 아닌 번안 검사가 중요함을 시사하고 있다. 실제로 한국과 미국간에 우울 또는 직장내 피로도의 기제 및 유발 요인이 전적으로 동일한 양상을 보이지 않을 것이다. 따라서 원저자로부터 권한을 인가 받고, 번안작업을 할 때 구조와 점수부여 체계를 유지하되 한국문화에 적절한 방향으로의 문항수정은 당연한 것이다. 또한 계약시에 부적절한 문항의 제거, 새로운 문항 추가 가능성을 명시해야 하는 것도 이러한 이유들 때문이고 한국문화에서 타당화 증거가 확보된 이후에 번안검사의 국내 사용이 가능하다.

[65] 공분산분석은 종속변수와 공변하는(상관이나 공분산들이 유의함) 변수 즉 공변수(covariate)들을 통제하고서 시행되는 분산분석이고(공분산 구조 분석과 다름), 블록내 무선화(randommized-block) 설계는 실험설계 가운데 블록내에서만 무선화(randomization)가 유지되는 설계이다. 블록 간에는 무선화가 적용되지 못한다.

[66] 측정되는 구성개념의 척도상에서 집단 차이가 의미있는 차이라면, 측정 방법에 관련된

실무를 위한 제안. 이것은 매우 중요한 지침이고 많은 분석이 수행될 수 있다. 동등성 분석의 경우, 실무에 대하여 다음과 같은 제안이 있다.

- 표본의 크기가 충분한 경우, 원본언어판과 대상언어판 간의 구성개념 동등성에 대한 비교 연구를 수행한다. 이러한 분석을 원활하게 해 줄 응용 프로그램들은 많이 있다(Byrne, 2006 참조).
- 언어 및 문화 집단 간 검사 구조의 일치도를 파악하기 위해 탐색적 요인 분석이나(바람직하게는, "목표 회전"의 사용: 정해진 구조를 지향해서 요인 구조를 회전하는 것을 말함) 또는 확인적 요인 분석 및 가중 다차원 척도법을 실시한다. 이러한 연구는 큰 표본을(측정변수[67] 수효의 10배) 요구하므로 비교문화연구가 어려워 질 수 있다. 이런 비교문화연구 유형에 대한 좋은 예는 Byrne과 van de Vijver (2014)의 연구에서 볼 수 있다.
- 수렴 및 변별 타당도[68]를 구한다 (필수적으로 구성개념들 사이의 상관들이 유의한지를 파악하고 이러한 관계가 언어 및/또는 문화적 집단 간에 안정성을 가지는지 확인한다(van de Vijver & Tanzer, 1997 참조).

DIF 분석의 실무에 대하여 아래에 몇 가지 제안사항이 있다. [물론] 보다 더 정교한 통계 방법을 사용하기 위해, 연구자들이 DIF에 관한 전문 문헌을 읽어 볼 것을 권장한다.

- 표준적인 절차 중 하나를 사용하여 DIF 분석을 수행한다(만약 문항이 이분으로 채점된 경우 Mantel-Haenszel 분석이 가장 간단하다. 만약 문항이 다분적으로 채점된 경우에는

언어/문화/경제 등 기타 측면에서의 집단차이는 원치 않는 차이가 된다.

[67] 탐색적 요인분석에서는 대체로 문항이 측정변수가 되고 확인적 요인분석에서는 연구자의 모형가설에 따라 문항, 문항묶음, 또는 소검사 등이 측정변수가 된다.

[68] 한차례 얻은 자료가 있을 때 수렴/변별 타당도를 파악하는 방법은 다음과 같다. 즉, 우선 하나의 방법(예, 설문조사)으로 수집된 자료일 경우, 요인 간 상관이 1.0보다 작으면 요인 간 변별타당도가 있다고 할 수 있다. 그러나 그 상관을 보기전에 각 요인의 분산이 0과 다른지(유의한지)를 보는 것이 각 요인(구성개념)의 수렴타당도를 보는 것이다. 이것은 상이한 측정치들간의 관계(상관)로부터 하나의 요인으로 수렴하는 부분이 있음을 보고자 하는 것이다. 그러나 한가지 방법만으로 수집된 자료는 동일 방법의 효과 때문에 요인분산이나 상관이 실제보다 증가할 수 있는 문제를 안고 있다. 따라서 원칙적으로는 다 특질 다방법(MTMM, multitrait-multimethod)설계를 통해서 수렴 및 변별타당도를 보는 것을 추천된다.

일반화된 Mantel-Haenszel 분석을 사용할 수 있다)[69]. 다른 좀 더 복잡한 방법으로는 IRT-기반의 방법이 있다. 만약 표본의 크기가 좀 더 작을 경우 "델타 도표[70]"를 사용해서 잠재적으로 결함이 있는 문항을 선별할 수 있다. 조건적 비교 [조건적 통계치 간 p 값의 비교][71]도 또다른 가능한 방법 중 하나이다 (표본이 작을 때 결과를 비교하는 방법으로는, Muñiz, Hambleton, & Xing 의 2001년 논문을 참조).

검수-3 (11). 대상집단용 번안 검사의 규준, 신뢰도 및 타당도를 뒷받침하는 증거를 제공한다.

설명. 원본 검사의 규준[72], 타당도 증거와 신뢰도 증거는 다른 언어/문화로의 번안판에 자동적으로 적용될 수 없다. 그러므로 개발된 모든 새로운 판에 대한 경험적 타당도와 신뢰도 증거를 제시해야 한다. 검사로부터 도출된 추론을 지지하는 모든 종류의 경험적 증거를 검사 교본에 수록해야 한다. 타당도 증거의 원천이 되는 검사의 내용, 반응 과정, 내적 구조, 다른 변수와의 관계 그리고 검사 결과 (AERA, APA, NCME, 2014)와 같은 다섯가지에 특히 주목해야 한다[73]. 탐색적, 확인적 요인 분석, 구조방정식 모형, 다특질-다방법 분석[74]은 내적 구조에 기초한

---

[69] 문항이 이분이거나 다분이거나에 관계없이, 측정오차를 통제하면서 DIF분석을 하려면 구조방정식모형에 기반한 RFA가 편리하다.

[70] Muñiz, Hambleton, 및 Xing(2001, p.120)에 소개된 델타도표(delta plot)는 대응기준(통상은 구성개념 점수에 대한 대체 점수로서 검사점수를 사용)에 따라 정의된 두 대응집단 각각에서 구한 문항 정답률(p)을 표준정규분포상에서의 점수(normal deviate)로 바꾼 값을 델타라하고, 집단별 델타 값을 좌표값으로 하여 2차원 평면(예: y축-원본검사 집단, x축-대상 언어 집단)상에 점을 찍으면 델타 도표이다. 이 도표에서 원점을 지나면서 45°기울기를 가지는 기준선에서 멀리 떨어진 점에 대한 문항들이 DIF의 가능성이 있다.

[71] 구성개념의 측면에서 같은 수준에 있는 사람들로 대응된 집단들에서의 통계치는 일종의 조건적(conditional) 통계치인데, 어떤 문항에 대한 이 통계치가 집단간에 차이가 크면 DIF의 가능성이 있다고 본다.

[72] 절대 평가일 경우(예, 다양한 자격 시험들, 임상 진단용 검사들), 규준(상대평가용임)이 아닌 기준점수가 작성된다. 전반적으로 이 지침서는 상대평가 중심으로 기술된 면이 있어, 절대 평가에 대한 내용들은 간간이 역자들이 각주에 제시하고 있다.

[73] 이들 다섯 가지 원천은 검사표준서(AERA, APA, NCME, 2014)에서 이야기하는 내용타당도, 구성개념 타당도, 준거 타당도, 결과 타당도를 위한 기본 정보가 된다.

[74] 다특질-다방법(MTMM: Multitrait-Multimethod)분석은 수렴타당도와 변별타당도를 보이는

타당도 증거 관련의 자료를 얻고 분석하는데 사용되는 통계적 기법이다.

실무를 위한 제안. 다음의 제안은, 사용을 고려중인 어떤 검사에서도 동일하게 요구되는 것들이다.

- 원본 검사의 규준(또는 기준점수)이 번안 검사에서도 적용되려면, 통계적으로 적절하고 [번안검사가] 공정하다[75]는 증거가 제공돼야 한다. 원본 검사의 규준(또는 기준점수)을 적용하는 것이 적절하다는 증거를 제공할 수 없는 경우, [검사표준서에 있는] 규준(또는 기준점수) 작성의 표준에 따라서 번안검사용 규준(또는 기준점수)을 별도로 개발해야 한다.
- 대상언어판 검사의 사용을 정당화할 수 있도록 신뢰도에 대한 충분한 증거를 확보한다. 통상적으로 내적 일관성의 추정치(예: KR-20 또는 알파계수 혹은 오메가 계수)를 제시할 수 있다[76].
- 번안 검사를 사용해야 할지 어떨지를 결정하는데 필요한 정도의 타당도 증거를 확보해야한다. 수집된 증거의 유형은 검사 점수가 어떻게 사용될 것인가에 따라 다를 것이다(예: 성취도 검사에는 구성개념 타당도의 증거가 필요하고, 적성검사에는 예측 타당도 등의 증거가 필요하다).

검수-4(12). 하나의 검사에 대한 여러가지 언어판 간에 점수체계를 연계할 경우 적절한 동등화(equating) 설계 및 자료 분석 절차를 사용해야 한다.

설명. 하나의 검사에 대한 두가지 언어판을 연계하여 단일한 점수체계를 만드는데, 몇 가지의 대안이 있다. 공통 문항군이 사용되는 경우, 두 언어 집단 간에 공통 문항군의 기능을 평가해야 하며, 만약 집단에 따라 차별적으로 기능하는 문항이 있으면, 두 언어판의 연계 설정에 필요한 공통문항군에서 제거할 것이 고려된다. 이런 목적으로 델타 도표(Angoff & Modu, 1973)가

방법인데 1960년대부터 사용되기 시작해서, 구성개념 타당화의 가장 강력한 접근 중 하나이다. 처음에는 상관행렬에 대하여 측정오차의 통제가 없이 MTMM 분석이 되었으나, 1980년대 중반에 구조방정식 모형의 소프트웨어가 보편화 되면서 측정오차를 통제하는 가운데 MTMM 분석이 가능해졌다. MTMM 자료에 대해 일반적인 공분산구조로 분석 후, 최종해를 표준화 해로 바꾸면 요인 간 또는 측정오차 간 상관을 해석할 수 있다.

[75] 검사의 공정성은 통계적 적절성과 다른 또 하나의 중요 영역이다.

[76] 측정의 표준오차도 함께 제시한다.

좋은데 Cook과 Schmitt-Cascallar(2005)는 집단 간 의미 차이가 있는 문항들을 식별하기 위해 델타도표를 어떻게 활용하는지를 보였다. 두 언어판간에 연계를 하고자 할 때 모든 유형의 문항들이 같은 정도로 유용한 것은 아니다. 공통 문항군에 대해 문항반응이론을 이용하면, 문항의 난이도 및 변별도 모수의 추정치를 도표로 작성하여 부적합한 문항을 식별할 수 있다(Hambleton et al., 1991 참조).

그런데 하나의 검사에 대한 두 가지 언어판 간에 점수를 연계(또는 "동등화")하는 것은 자료에 대한 강한 가정이 필요하기 때문에 많은 어려움이 있다. 때로는 한 검사의 여러 언어판들이 [내용상] 동등해서 검사 점수를 상호교환적으로 사용할 수 있다고 하는 가정이 큰 문제를 일으키기도 한다. 수학시험이라면 번역/번안이 간단해서 그런 가정이 무난할 수 있다. 만약에 하나의 검사에서 두 가지 언어 판이 세심하게 제작된 경우에는, 원본언어판이 그 언어 집단에서 기능하는 것과 동등하게 번안판이 대상언어 집단에서 기능한다는 가정을 할 수 있다. 그런데 이런 가정은 해당 검사의 두 언어판이 동등하고 대상언어판의 점수에 영향을 미치는 방법 편향이 없는 경우에 가능하다.

[점수 연계를 위한] 두 가지 대안이 있지만, 어느 것도 완벽하지는 않다. 첫번째 방법으로는, 해당 검사의 두 언어판에서 본질적으로 동등하다고 간주되는 문항군에 대해서 연계를 수행할 수 있다. 예를 들어, 그 문항들이 매우 쉽게 번역/번안될 수 있다고 판단된 경우이다. 원칙적으로 이 대안은 가능한 방법이지만 연계에 사용되는 문항군과 나머지 문항군 간에 구성개념이 같아야 한다. 두 번째 대안은 이중 언어자들을 통해 연계를 수행하는 것이다. 이들이 두 가지 판에 응답한 결과를 가지고 점수 변환표를 작성하는 것이 가능하다. 이때 표본의 크기가 너무 작으면 안 되며, 설계에서 [각 언어판] 검사 양식의 제시 순서는 서로 균형[77]을 이루어야 한다. 이 방법에서 강력한 가정은 참가자들이 이중 언어를 능통하게 사용할 수 있어서, 난이도 차이를 통제하면 참가자는 두 언어판 간에 동일한 수행을 보인다는 것이다. [이 때] 두 언어판 간에 어떤 차이가 있다면, 그것은 한 언어판에서 다른 언어판으로 점수를 변환할 때 개별 점수의 수정에 사용된다[78].

실무를 위한 제안. 어떠한 동등화 방법들도 최소한 하나 이상의 단점이 있기 때문에 번안판 간에 점수를 연계하는 데는 어려움이 있다. 아마도 가장 좋은 전략은 점수의 동등화를 확립하기 위한 모든 단계를 완벽하게 따르는 것이다. 아래의 세 가지 질문에 대응할 수 있는

---

[77] 서로 균형(counterbalance)이란, 예로서 영어판과 한글판을 이중언어자 표본에게 실시할 때 영어-한글의 순서로 실시되는 사람수와 한글-영어의 순서로 실시되는 사람수가 같음을 의미한다

[78] 한 언어판 검사의 평균이 5.5이고 다른 언어판 검사의 평균이 5이면 평균차이가 0.5이고 최종적으로 다른 언어판을 수행한 참가자는 원점수에 0.5를 더한 변환점수가 최종 수행점수가 된다.

증거가 강력하면 한 검사에 대한 두 가지 언어판 간에 검사 점수를 서로 상호 교환적으로 사용할 수 있다.

- 검사의 원본과 대상언어 판에서 동일한 구성개념이 측정되고 있다는 증거가 있는가? 그 구성개념이 새로운 문화에서, 다른 외부 변수와 가지는 관계가 원본언어 집단에서와 같은가?
- 방법 편향의 원인이 제거되었다는 강력한 증거가 있는가(예, 시간 관련 문제없음, 검사에 사용된 형식이 집단 간에 동등하게 친숙함, 지시에 대한 혼란이 없음, 각각의 응답자 집단이 모집단에 대해서 가지는 대표성에 체계적인 문제가 없음, 표준화된 지시사용, 동기부여 수준에서의 차이나 극단적 선지의 선택과 같은 특정 반응 경향이 없음)?
- 검사에 편향가능성 있는 문항은 없는가? 여기서, 두가지 판 간에 문항의 p 값, 더 바람직하게는 델타값이 매우 유용할 수 있다. [델타 도표에서] 원점을 지나는 45°선 위에서 멀리 떨어지는 문항이 두 언어간에 동등하게 작용하는지의 여부를 결정하기 위해 살펴 보아야 한다. DIF 분석은 언어와 문화 집단 간 동등성에 대해서 훨씬 더 강력한 증거를 제공한다.
- 점수의 연계화를 시도한다면, 적절한 연계 설계 모형을 선택하고 구현할 필요가 있다. 그리고 연계 설계의 타당도에 대한 증거들을 제공해야 한다.

4. 실시 지침

실시-1 (13). 실시절차로 인해 발생하는 언어/문화적 문제 및 점수의 타당도를 저해할 수 있는 반응방식(response mode)이 최소화될 수 있는 실시자료와 지시사항을 준비한다.

설명. 실시 지침의 이행은 특정 언어/문화 맥락에서 검사 점수의 타당도를 위협할 수 있는 모든 요소에 대한 분석부터 시작해야 한다. 단일한 언어/문화 맥락에서의 실시 경험은 다중적인 언어/문화에서 발생할 수 있는 문제를 예측하는 데 도움이 될 수 있다. 예를 들어, 숙련된 실시자는 지시사항의 어떤 측면이 응답자에게 어려운지 알 수가 있다. 이러한 측면들은 번역이나 번안 후에도 여전히 어려울 수 있다. 새로운 언어적 또는 문화적 맥락에서 검사도구를 사용하는 것은 이전 단일 문화에서는 발견되지 않은 다른 문제를 가져올 수가 있다.

실무를 위한 제안. 이 지침에서는 검사 실시에 문제를 일으킬 수 있는 잠재적 요소들을 예측하는 것이 중요하다. 검사실시의 공정성을 확보하기 위해 연구해야 할 몇 가지 요소들은 다음과 같다.

- 검사 지시사항의 명확성(지시사항의 번역 포함), 응답 기구(예, 답안지), 허용 시간(예,

응답자가 마칠 수 있는 충분한 시간을 주지 못하는 것은 일반적 오류의 원인이 된다), 응답자가 검사를 완료할 수 있도록 동기부여, 검사 목적에 대한 지식, 그리고 채점방식.

실시-2 (14). 모든 대상집단에서 엄격하게 준수해야 하는 검사 실시의 요건들을 명시한다.

설명. 이 지침의 목적은 검사 개발자가 모든 대상집단에서 엄격하게 준수해야 하는 검사 지시사항 및 관련 절차(예, 검사실시의 요건, 시간 제한 등)를 확립하도록 고무하는 것이다. 이 지침은 주로 검사 실시자가 표준화된 지침을 따르게 하기 위한 것이다. 동시에, 추가 시간, 확대된 글씨, 특별히 저소음의 실시 환경 등을 적용해야 하는 하위 집단에 대한 양해사항을 명시하게 한다. 오늘날 검사 분야에서는 이러한 것을 "검사 양해사항"이라고 한다. 이러한 검사 양해사항의 목표는 응시자의 점수를 높이자는 것이 아니라, 그들이 느끼고 알고 할 수 있는 것을 그대로 보여줄 수 있는 검사 환경을 조성하는 것에 있다.

표준화된 검사 요건에 변이(variation)가 도입될 경우 그에 주목하고, 다음에 그 변이 및 그것이 [결과의] 해석에 미치는 영향을 고려할 수 있도록 해야 한다.

실무를 위한 제안. 본 지침은 부분적으로 실시-1 (13)과 중복될 수 있지만, 응답자들이 가능한 한 유사한 조건에서 검사를 받는 것이 중요함을 강조하기 위해 반복해서 언급되었다. 두 언어판의 점수가 상호교환적으로 사용될 경우 이 지침은 필수적인데 다음과 같은 몇 가지 제안이 있다.

- 검사 지시사항과 관련 절차는 새로운 언어와 문화에 적합한 표준화된 방법으로 번안되어야 하고 재작성되어야 한다.
- 검사 지시 사항과 관련 절차를 새로운 문화에 맞게 변경할 경우, 실시자를 새로운 절차에 맞추어 훈련시키고, 원본 검사의 절차가 아닌 새로운 절차를 존중하도록 지도한다.

5. 점수체계 및 결과해석 지침

점수체계 및 결과해석-1 (15). 집단 간 점수 차이를 해석할 때는 가용한 모든 정보를 참조한다.

설명. 비록 검사가 기술적으로 탄탄한 절차를 거쳐 번안되었고 검사 점수의 타당도가 어느 정도 확립되었다 하더라도, 검사 응답자들의 국적 및/또는 문화가 다르면 문화 차이 및 기타 여러 차이가 발생하므로 집단 간 차이의 의미가 여러 면에서 해석될 수 있음을 명심해야 한다. Sireci (2005)는 같은 언어/문화 집단에 속한 이중 언어 사용자들을 응답자로 하여, 한 검사에 대한

두 가지 언어판을 실시하고 동등성을 평가하는 방식들을 검토하였다. 그는 이중 언어를 구사하는 응답자를 이용하여 할 수 있는 두 언어판 간 동등성 연구에 대한 몇 가지 연구방안의 개요를 설명하고, 통제해야 할 오염변수를 제시하였으며, 조사 결과를 해석하는 가치 있는 몇 가지 제안을 하였다.

실무를 위한 제안. 실무를 개선하기 위한 하나의 제안은 다음과 같다.

- 연구 질문(또는 집단 간 비교가 이루어지는 맥락)에 따라, 최종적으로 하나의 해석을 결정하기 전에 가능한 많은 해석을 생각해 볼 수 있다. 예를 들어, 한 집단이 다른 집단보다 점수가 높다고 추론하기 전에 검사를 잘 수행하려는 동기의 차이를 통제하는 것이 중요하다. 또한 검사 수행에 상당한 효과를 미치는 배경적 영향도 있을 수 있다. 예를 들어, 한 집단의 사람들은 효과적이지 못한 교육 시스템 가운데 있을 수가 있고, 이것은 검사 수행에도 상당한 영향을 미칠 것이다.

점수체계 및 결과해석-2 (16). 보고되는 점수체계의 동일성이 확립[79] 된 경우에만 집단 간 점수를 비교한다.

설명. 번역 및 번안의 핵심 목적이 언어/문화 집단 간 비교 연구인 경우, 해당 검사에 대한 복수의 언어판들은 공통의 점수체계를 사용해야 한다. 이는 "연계" 또는 "동등화"로 불리는 과정을 통해 수행된다. 이 과정에서 상당한 크기의 표본이 필요하며, 해당 검사의 번안판들간에 구성개념, 측정방법, 및 문항 측면에서 편향이 없다는 증거가 필요하다.

Van de Vijver와 Poortinga (2005)는 언어/문화 집단 간 검사 동등성의 몇가지 수준을 분류하였고 이들의 연구는 검사동등성 개념을 이해하는 데 충분한 도움이 된다. 사실, 검사동등화에 대한 창의적 개념이 그 두 저자에 의해 소개되었다. 측정 단위의 동일성은 각 집단의 점수체계에서 단위크기(magnitude)[80] 가 같다는 것으로, 집단 내 개인점수의 차이가 두집단에 걸쳐 같은 의미를 갖는다는 뜻이다 (이럴 경우, 예로서, 중국 표본에서 남녀 간 점수 차이를 프랑스 표본에서 남녀 간 점수 차이와 비교할 수 있다). 그러나 집단 간 점수의 직접

---

[79] 다집단 분석에 의해서 집단간 측정체계(예, 측정틀, 측정원점, 측정단위, 측정오차 분산 등)의 동일성이 검증되었음을 의미한다. 측정체계가 동일하면 집단간 검사의 동등성에 대한 양적 근거가 된다.

[80] 단위크기란 구성개념 1단위에 대하여 부여되는 실제 점수값을 말한다. 요인분석의 경우 요인점수에서 1의 차이가 날 때 측정변수 점수에서 나타나는 점수 차이인데 통계적으로는 요인계수(factor loading)가 된다.

비교가 타당해지려면 원점 동일성[81] 또는 검사점수 동일성이라는 가장 높은 수준의 동일성이 성립해야 한다. 이러한 가장 강한 동일성의 조건은 각 집단 간에 측정 단위가 같고 측정 원점이 같아야 한다.

한 검사에 대한 두 언어판 간 점수의 연계 또는 동등화를 위한 수많은 방법(고전검사이론과 문항반응이론의 측면 모두에서)이 제시되었다. 관심있는 독자들은 이 주제에 대한 더 깊은 이해를 얻기 위해 Angoff (1984) 및 Kolen과 Brennan (2004)의 책을 참조하길 바란다. Cook과 Schmitt-Cascallar (2005)는 현재 교육 및 심리 검사 척도의 동등화 및 점수체계 설정(Scaling)에 사용 가능한 통계 방법들을 이해하는데 필요한 기초를 제공했다. 저자들은 검사 번안 연구에 사용된 점수연계 절차를 기술하고 비평했으며, 지난 20년간 미국의 학업능력평가시험(SAT) 점수를 스페인어판 학업 적성검사 점수와 연계시키기 위해 수행된 3개의 연구를 기술하고 비평하는 가운데 사용된 연계 절차와 문제점을 선별적으로 제시하였다.

실무를 위한 제안. 여기서 핵심은 검사 점수를 과대 해석하면 안 된다는 것이다.

- 가능한 타당도 증거의 범위내에서 결과를 해석해야 한다. 예를 들어, 두 언어 집단 간 측정 동일성이 확립되지 않는 한 검사점수 수준을 비교하는 어떤 진술도 하지 않는다.

6. 문서화 지침

문서화-1 (17). 검사가 다른 집단에서 사용될 수 있도록 번안될 때는, [원본 검사와의] 동등성을 지지하는 모든 증거에 대한 설명을 포함하여, 모든 변경사항을 기술 교본에 기록해서 제공한다[82].

설명. 본 지침의 중요성은 많은 연구자에 의해 인식되고 강조되고 있다(예를 들어, Grisay,

---

[81] 측정 원점 동일성(Scalar invariance)에서 원점은 구성개념(F)을 측정치(X)에 의해 측정하는 과정을 나타내는 1차식(X=a+bF+e)에서 절편(a)을 의미하는데, 구성개념이 0일 때 부여되는 측정치의 값이므로 측정원점(origin)이라고도 한다. 위의 1차식에서 b는 측정단위이다. 통상적으로 측정단위 동일성이 성립해야 측정원점 동일성을 검증하게 된다. 만일에 집단간에 단위도 원점도 동일하면 측정치(실제는 검사점수)가 동일하다고 해석된다.

[82] 검사제작후에 두가지 종류의 교본이 작성된다. 하나는 사용자들을 위해 검사의 목적 및 유용성, 그리고 실제 사용의 실무적 사항을 중심으로 한 사용자 교본이고, 나머지는 검사의 품질을 검토하는 검사이론의 전문가들을 위해 기술적 사항을 총괄해서 기록한 기술 교본(통상 개발 보고서의 형식을 취함)이다. 따라서 번안 검사에도 사용자 교본에 더해서 기술교본이 제공된다. 사용자 교본은 문서화-2(18)에서 설명된다

2003 참조). TIMSS 와 PISA는 번안 작업 전반에 걸쳐 변경사항을 상세하게 문서로 만들어 이 지침을 성공적으로 준수해 왔다. 이러한 문서화는, 이후에 도입되는 변경사항이 적절한지의 판단을 위한 근거가 된다.

또한 기술 교본에는, 나중에 연구자들이 동일 집단 또는 다른 집단에 그 절차를 재현할 수 있도록 방법론에 대하여 충분히 상세한 정보가 수록되어야 한다. 기술교본에는 새로운 집단에서 번안된 검사의 사용을 지지하는 구성개념 동등성, 점수체계 동등성([연계]가 실행된 경우)의 증거에 대한 충분한 정보가 수록되어야 한다. 집단 간 비교가 이루어질 경우, 기술 교본에는 집단 간 점수의 동등화를 결정하는데 사용된 증거를 보고 해야 한다.

때로는, 기술 교본에 관심있어 하는 독자로부터 질문이 제기될 수 있다. 이 교본은 이 분야 전문가는 물론, [검사가] 새로운 다른 집단에 사용되기 전에 그 검사의 유용성을 평가하는 사람들을 위해 작성되어야 한다 (비전문가의 이해를 위해 간단한 보충 문서가 추가될 수 있다).

실무를 위한 제안. 번안 검사에는 번안 과정과 관련된 모든 양적, 질적 자료들을 문서화한 기술 교본이 있어야 한다. 특히 번안 과정에서 제2의 언어 및 문화권에 맞게 변경된 사항을 모두 문서화하는 것이 도움이 된다. 기본적으로 이 분야 전문가와 학술지 편집자는 검사의 번안판을 개발하고 타당화하는데 투입된 모든 과정에 대하여 기술된 문서를 원할 것이다. 물론, 그들은 모든 분석 결과를 보고 싶어 할 것이다. 다음은 그들에게 답변해야 할 질문의 유형이다.

- [검사에서 측정되는] 구성개념 및 새로운 집단에 맞춰 번안된 검사의 효용성을 지지하는 증거는 무엇인가[83]?
- 어떠한 표본에서 어떠한 문항들에 대한 자료가 수집되었는가[84]?
- 내용타당도, 준거타당도, 그리고 구성개념 타당도를 평가하기 위해 어떤 다른 자료를 구했는가[85]?
- 다양한 자료들이 어떻게 분석되었는가?

---

[83] 번안검사의 평가자나 소비자라면 우선 해당 구성개념의 조작적 정의 또는 그것을 측정할 필요에 대한 증거(구성개념의 효용성)를 찾고자 할 것이다. 그리고 나면 그것을 측정하는 번안검사의 타당도 증거(번안검사의 효용성)를 보고자 할 것이다.

[84] 문항관련 자료이므로 예비검사 단계에서의 내용이 중심이 된다.

[85] 타당화단계이므로 본검사 및 필요에 따른 추가검사 단계를 의미한다.

- 그 결과물들은 무엇인가?

문서화-2 (18). 새로운 집단에서 번안 검사를 사용하는 실무자를 지원해줄 사용자 교본을 제공한다.

설명. 이 문서는 실제 평가장면에서 번안 검사를 사용할 사람을 위해 작성되어야 한다. 검사 이용에 관한 국제검사위원회(ITC) 지침서에서 정한 모범적 실무와 일치해야 한다(www.InTestCom.org 참조).

실무를 위한 제안. 검사 개발자는 집단의 사회 문화 및 생태학적 환경이 검사 수행에 영향을 미칠 수 있는 가능성에 대하여 구체적인 정보를 제공해야 한다. 사용자 교본은 다음과 같이 구체적으로 구성되어야 한다.

- 검사에 의해 측정되는 구성개념을 설명하고 정보를 요약한다. 즉 번안 과정을 기술한다.
- 문항 내용의 문화적 적절성, 검사 지시사항, 반응 형식 등 번안을 뒷받침하는 모든 증거를 정리한다.
- 대상집단 내의 다양한 하위집단들에 걸친 검사 사용의 적절성 및 기타 사용상의 제약 조건에 대하여 정의한다.
- 검사 실행의 모범적 실무에 관련하여 검토해야 할 사항들에 관해 설명한다.
- 집단 간 비교가 가능한지, 어떻게 비교할 수 있는지를 설명해야 한다.
- 채점과 규준(예, 관련 규준표)[86]을 위해 필요한 정보를 제공하거나 사용자가 어떻게 채점 절차(예, 컴퓨터 기반 검사)에 접근할 수 있는지 설명한다.
- 검사 점수로부터의 추론에 관한 타당도 및 신뢰도 자료가 가지는 의미에 대한 정보를 포함하여 결과 해석에 관한 지침들을 제공한다.

맺음말

우리는 [번안] 검사 개발자와 사용자들의 업무 수행을 돕기 위한 일련의 지침을 전달하기 위해

[86] 검사의 목적이 규준참조검사(상대평가)일 때는 규준이 있을 것이고 준거(영역)참조 검사(절대평가)라면 기준점수가 있을 것이다.

최선을 다했다. 그러나 이러한 지침들 그리고 부실한 관행을 바꾸려는 노력이 효과를 거두기 위해서는 이를 전파하는 좋은 장치가 있어야 한다. Rios와 Sireci(2014)가 최근 체계적으로 검토한 결과, 문헌에 보고된 대부분의 검사 번안이, 이미 20년의 역사를 가진 ITC 지침을 따르지 않았다. 따라서 우리는 독자들이 동료들에게, 전 세계의 많은 전문가들이 기여한 최선의 실무 참고서로서의 지침서 2판에 대한 인지도를 높이기 위해 모든 노력을 기울일 것을 권한다.

동시에, 우리는 이 지침의 첫 번째 판이 현재 바뀌는 것처럼, 이어서 이번 두 번째 판의 지침들도 언젠가는 바뀔 것이다. 잘 알려진 AERA, APA, NCME 검사 표준서는 현재 6판(AERA, APA, & NCME, 2014)까지 나와 있다. 우리는 번안검사에 대한 ITC의 지침도 앞으로 또 다른 개정을 겪을 것으로 기대한다. 인용될 필요가 있거나 제3판에 영향을 미치는 새로운 연구를 알고 있는 경우, 또는 여기에 제시된 18개 지침에 대한 새로운 지침이나 수정을 제공하고자 하는 경우 ITC에 알려주기 바란다 (www.InTestCom.org). 제2판을 제작한 현 "연구/지침 위원회" 위원장 또는 www.InTestCom.org 에 있는 이메일 주소로 ITC 간사에게 연락하면 된다.

감사의 인사말

국제검사위원회(ITC)는 몇 년 동안 번안검사 지침서 제2판을 만들기 위해 일한 6명의 위원들에게 감사를 표하고자 한다. Dave Bartram, SHL, UK; Giray Berberoglu, Middle East Technical University, Turkey; Jacques Grégoire, Catholic University of Louvain, Belgium; Ronald Hambleton, Committee Chairperson, University of Massachusetts Amherst, USA; Jose Muñiz, University of Oviedo, Spain; and Fons van de Vijver, University of Tilburg, Netherlands.

또한 국제검사위원회는 Chad Buckendahl (미국), 영국에 있는 Anne Herrmann 과 그녀의 동료들 OPP Ltd. (영국), 그리고 매사추세츠 대학교(미국)에 있는 April Zenisky, 및 켄트 대학교(영국)에 있는 Anna Brown 에게 이 문서의 초안을 주의 깊게 검토해준 것에 대해 감사를 표한다. ITC는, 직간접적으로 ITC의 번역/번안 검사 지침서 제2판에 공헌한 전 세계의 다른 모든 검토자들에게도 감사드린다.

참고문헌

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Modu, C. C. (1973). *Equating the scales of the Prueba de Apititud Academica and the Scholastic Aptitude Test (Research Rep No. 3).* New York: College Entrance Examination Board.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural modeling. *Structural Equation Modeling, 16*, 397-438.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 137-164). Newbury Park, CA: Sage Publications.

Byrne, B. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing, 1*, 55-86.

Byrne, B. (2003). Measuring self-concept measurement across culture: Issues, caveats, and application. In H. W. Marsh, R. Craven, & D. M. McInerney (Eds.), International advances in self research(pp. 30-41). Greenwich, CT: Information Age Publishing.

Byrne, B. (2006*). Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Publishers.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psichothema, 20*, 872-882.

Byrne, B. M., & van de Vijver, F.J.R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10*, 107-132.

Byrne, B. M., & van de Vijver, F.J.R. (2014). Factorial structure of the Family Values Scale from a multilevel-multicultural perspective*. International Journal of Testing, 14*, 168-192.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripley, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational*

Measurement, 33(2), 202-214.

Cook, L. L., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 139-170).

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and Practice* (pp. 137-166).

Ellis, B. B. (1989). Differential item functioning: Implications for test translation. *Journal of Applied Psychology, 74*, 912-921.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77*, 177-184.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*(6), 543-533.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3), 199-215.

Ercikan, K., Gierl, J. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301-321.

Ercikan, K., Simon, M., & Oliveri, M. E. (2013). Score comparability of multiple language versions of assessments within jurisdictions. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *An international handbook for large-scale assessments* (pp. 110-124). New York:

Grégoire, J., & Hambleton, R. K. (Eds.). (2009). Advances in test adaptation research [Special Issue]. *International Journal of Testing, 9(*2), 73-166.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225-240.

Hambleton, R. K. (2002). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, NJ: Lawrence Erlbaum Publishers.

Hambleton, R. K., & de Jong, J. (Eds.). (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20*(2), 127-240.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Applied Testing Technology, 1*(1), 1-16.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*(1), 1-18.

Hambleton, R. K., & Lee, M. (2013). Methods of translating and adapting tests to increase crosslanguage validity. In D. Saklofske, C. Reynolds, & V. Schwean (Eds.), *The Oxford handbook of child assessment* (pp. 172-181). New York: Oxford University Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Hambleton, R. K., Yu, L., & Slater, S. C. (1999). Field-test of ITC guidelines for adapting psychological tests. *European Journal of Psychological Assessment, 15* (3), 270-276.

Hambleton, R. K., & Zenisky, A. (2010). Translating and adapting tests for cross-cultural assessment. In D. Matsumoto & F. van de Vijver (Eds.), *Cross-cultural research methods* (pp. 46-74). New York, NY; Cambridge University Press.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Javaras, K. N., & Ripley, B. D. (2007). An 'unfolding' latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association, 102*, 454-463.

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission Guidelines: Keeping validity in mind. *European Journal of Psychological Assessment, 15*(3), 277-283.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika, 68*, 563-583.

Kolen, M. J., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Levin, K., Willis, G. B., Forsyth, B. H., Norberg, A., Stapleton Kudela, M., Stark, D., & Thompson, F. E. (2009). Using cognitive interviews to evaluate the Spanish-language translation of a dietary questionnaire. *Survey Research Methods, 3*(1), 13-25.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*(2), 115-135.

Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traduccion y adaptacion de los tests: segunda edicion. *Psicothema, 25*(2), 149-155.

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 1*(2), 115-135.

Oort, F. J., & Berberoğlu, G. (1992). Using restricted factor analysis with binary data for item bias detection and item analysis. In T. J. Plomp, J. M. Pieters, & A. Feteris (Eds.), *European Conference on Educational Research: Book of Summaries* (pp. 708-710). Twente, the Netherlands: University of Twente, Department of Education.

Park, H., Pearson, P. D., & Reckase, M. D. (2005). Assessing the effect of cohort, gender, and race on DIF in an adaptive test designed for multi-age groups. *Reading Psychology, 26,* 81-101.

Rios, J., & Sireci, S. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing, 14*(4), 289-312.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and MantelHaenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Rotter, J.B. & Rafferty, J.E. (1950). *Manual: The Rotter Incomplete Sentences Blank*: College Form. New York: Psychological Corporation.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and Black examinees. *Applied Measurement in Education, 10*(4), 299-319.

Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice, 16*, 12-19.

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). Mahwah, NJ: Lawrence Erlbaum Publishers.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 148-166.

Sireci, S. G., & Berberoğlu, G. (2000). Using bilingual respondents to evaluate translatedadapted items. *Applied Measurement in Education, 13*(3), 229-248.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. Merenda, & C. Spielberger, C. (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-116). Mahwah, NJ: Lawrence Erlbaum Publishers.

Sireci, S. G., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats. *International Journal of Testing, 3*(2), 129-150.

Sireci, S. G., & Wells. C. S. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 33-68). Washington, DC: Council of Chief State School Officers.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2*(2), 107-129.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic

regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Tanzer, N. K., & Sim, C. O. E. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC Guidelines for Test Adaptation. *European Journal of Psychological Assessment, 15*, 258-269.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147169). Mahwah, NJ: Lawrence Erlbaum Publishers.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 67-113). Mahwah, NJ: Lawrence Erlbaum Publishers.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89-99.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage Publications.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31*, 33-51.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277-308). Dordrecht, the Netherlands: Kluwer Academic Publishers.

van de Vijver, F. J. R., & Poortinga, Y. H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment, 8,* 17-24.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in crosscultural assessment. *European Journal of Psychological Assessment, 13*, 29-37.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodical issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-64). Mahwah, NJ: Lawrence Erlbaum Publishers.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*(4), 263-279.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913-934.

부록 A.

검사 번역 및 번안을 위한 ITC 지침 점검표

여기에는 18개의 ITC 지침을 상기시키기 위한 점검표가 있다. 우리는 당신의 검사 번역/번안 과제에서 만족스럽게 처리했다고 느끼는 것들을 체크하고, 또한 처리되지 않은 것들에 주의를 기울일 것을 권고한다.

선행조건 지침

- ▢ 선행조건-1 (1).
  번안하기 전에, 검사와 관련된 지적 재산권 소유자로부터 허가를 얻는다.
- ▢ 선행조건-2 (2).
  검사점수를 적용하고자 하는 대상 집단에 비추어, 검사에서 측정될 구성개념의 정의와 내용이 문항내용에 반영된 정도가 충분한지 평가한다.
- ▢ 선행조건-3 (3).
  대상집단에서 검사 의도와 관련 없는 언어/문화적 차이가 가져오는 영향을 최소화한다.

검사 개발 지침

- ▢ 검사개발-1 (4).
  번역 및 번안 과정에서 관련 전문 지식을 갖춘 전문가를 택하여, 대상집단의 언어적, 심리적, 문화적 차이에 대한 고려를 확실히 한다.
- ▢ 검사개발-2 (5).
  대상집단에서 검사 번안의 적합성을 극대화하기 위해 적절한 번역 설계와 절차를 사용한다.
- ▢ 검사개발-3 (6).
  검사에서의 지시사항과 문항 내용이 모든 대상집단들에서 유사한 의미를 갖는다는 증거를 제공한다.
- ▢ 검사개발-4 (7).
  문항 형식, 평정 척도, 점수부여 범주, 검사상의 관례, 실시 방법 및 기타 절차가 의도된 모든 대상집단에 적절하다는 증거를 제공한다.
- ▢ 검사개발-5 (8).
  문항분석, 신뢰도 평가 및 소규모의 타당도 연구를 통해 번안검사에 수정할 수 있도록 예비검사를 실시하여 자료를 수집한다.

검수 지침

- ▢ 검수-1 (9).
  검사의 대상집단으로서의 특징을 가지며 경험적 분석을 위한 충분한 크기의 표본을

선정한다.

- 검수-2 (10).
  집단 간에 구성개념의 동등성, [측정] 방법 동등성 및 문항 동등성에 대한 관련 통계 수치를 제공해야 한다.
- 검수-3 (11).
  대상집단용 번안 검사의 규준, 신뢰도 및 타당도를 뒷받침하는 증거를 제공한다.
- 검수-4 (12).
  하나의 검사에 대한 여러가지 언어판 간에 점수체계를 연계할 경우 적절한 동등화(equating) 설계 및 자료 분석 절차를 사용해야 한다.

실시 지침

- 실시-1 (13).
  실시절차로 인해 발생하는 언어/문화적 문제 및 점수의 타당도를 저해할 수 있는 반응방식(response mode)이 최소화될 수 있는 실시자료와 지시사항을 준비한다.
- 실시-2 (14).
  모든 대상집단에서 엄격하게 준수해야 하는 검사 실시의 요건들을 명시한다.

점수체계 및 결과해석 지침

- 점수체계 및 결과해석-1 (15).
  집단 간 점수차이를 해석할 때는 가용한 모든 정보를 참조한다.
- 점수체계 및 결과해석-2 (16).
  보고되는 점수체계의 동일성이 확립된 경우에만 집단 간 점수를 비교한다.

문서화 지침

- 문서화-1 (17).
  검사가 다른 집단에서 사용될 수 있도록 번안될 때는, [원본검사와의] 동등성을 지지하는 모든 증거에 대한 설명을 포함해서, 모든 변경사항을 기술 교본에 기록해서 제공한다.
- 문서화-2 (18).
  새로운 집단에서 번안 검사를 사용하는 실무자를 지원해줄 사용자 교본을 제공한다.

부록 B.

용어집

가중된 다차원 척도법(WMDS; Weighted Multi-Dimensional Scaling). 이것은 차원성 검증을 다루는 또 다른 통계적 절차이다.

검사 점수 동등화. 동일한 구성개념을 측정하지만 검사가 엄격하게 평행[87]하지 않은 두 검사에서 점수를 연계하는 통계적 절차.

검사의 동시 개발. 번역 품질의 통제를 위한 표준화 절차를 사용하여, 원본언어판과 대상언어판을 동시에 개발한다. 대규모의 국제적 [검사] 프로젝트는, 한 언어로 개발된 검사가 나중에 다른 언어로 번역/번안될 수 없게 되는 경우를 피하기 위해 처음부터 여러 언어판을 동시에 개발하려는 경향이 증가하고 있다.

검사의 차원성. 검사가 측정하는 차원 또는 요인의 수를 말한다. 이에 대한 분석은 고윳값 도표나 구조방정식 모형을 포함한 많은 절차 중 하나를 사용하여 통계적으로 수행된다.

구조방정식 모형. 하나의 검사내 또는 [소검사들로 이루어진] 검사집에서의 내적 구조를 식별하는데 사용되는 복잡한 통계적 모형의 집합체이다. 이러한 모형들은 변수들 사이의 관계에 대한 인과 추론[88]을 하는 데 자주 사용된다.

국제학업성취도평가(PISA; Programme for International Student Assessment). 40개 이상의 참가국을 가진 경제협력개발기구(OECD)가 후원하는, 학업성취에 대한 국제적 평가이다.

대상언어. 하나의 검사를 번역/번안하는데 사용되는 언어. 예를 들어, 검사를 영어에서 스페인어로 번역하면 영어판을 '원본언어판"이라 부르고, 스페인어판을 "대상언어판"이라 부른다.

델타값. 델타값은 단순히 비선형적으로 변환된 z 값이며 이분 채점된 문항에 적용된다. 문항의 델타값은 문항의 *p*값을 표준정규분포 (평균 = 0.0, 표준편차 = 1.0)에서의 정규점수(normal

---

[87] 검사이론에서 평행의 의미는 두 검사간에 내용과 신뢰도가 같을 때를 의미한다. 검사 간에 절대평행이 가능하다면 두 검사 점수의 합산 또는 평균점수를 사용해야 할 것이지만 현실이 그렇지 않으므로 점수의 연계 즉, 검사 점수 동등화를 시도한다.

[88] 인과추론에는 몇가지 조건이 있다. 즉, 원인과 결과의 기제가 있을 것, 원인이 결과에 비해 시간적으로 우선함, 원인변수의 수준을 변화시키면 결과 변수에 변화를 가져올 것 등의 조건이 만족되어야 인과추론이 가능하다. 그러한 조건들의 확보없이 구조방정식 모형만으로는 단지 변수간 '함수적 관계'를 파악할 수 있을 뿐이다. 따라서 연구자는 구조방정식 모형검정에서 구한 '함수적 관계'라는 양적 정보에 질적 정보 및 이론적 개념을 더하여 인과추론을 수행한다.

deviate)로 바꾼 값인데, 그 값 아래의 영역(넓이)은 해당 문항에 정답을 한 응답자 비율(*p*)과 동일하다. 따라서 *p*가 0.84이면 문항의 델타값은 "1.0"이 된다. 이러한 변환은 [*p*값에 비해] 델타값이 등간 척도를 따를 가능성이 더 높다는 믿음에 기반한다.

두벌-번역 및 조정. 이 번역 설계에서는 [두벌의 순번역이 있고] 번역자 또는 전문가 패널이 두벌의 순번역간 차이를 식별하고, 그것들을 단일판으로 조정한다.

문항반응이론. 문항에 의해 측정되는 속성 또는 속성들에 반응을 연결하기 위한 통계적 모형이다. 구체적으로 IRT모형들은 이분 문항과 다분 문항의 자료를 모두 처리할 수 있다. 이분 반응자료는 성격척도에서 참/거짓으로 응답되는 문항을 채점하여 얻게 된다. 다분 반응자료는 성취도 평가에서 실기 과제나 작문을 채점할 때 또는 "리커트 척도"와 같은 평정척도에서 수집될 수 있다.

수험자. 검사 분야에서 "검사 수검자", "지원자", "응답자", 및 "학생(성취도 검사)"은 상호 교환적으로 사용된다[89].

수학 및 과학 연구의 국제적 동향(TIMMS; Trends in International Mathematics and Science Study). IEA가 후원하고, 수학 및 과학 분야에서 각 국가의 4, 8, 12학년 학생들을 대상으로 평가하는 국제적 평가를 말한다.

순방향 번역 [순번역] 설계. 이 설계에서는 번역자 또는 번역자 집단에서 검사를 대상언어로 번안한 다음, 다른 번역자 또는 번역 집단이 원본 검사와 번안검사 간 동등성을 판단한다.

알파[90] (혹은 "알파 계수" 또는 "크론바크 알파"라고도 불린다) 문항이 공통으로 하나의 속성을 측정하고 동일한 변별력을 가진다는[91] (따라서 오메가의 특수한 경우이다-아래 참조) 가정하에 계산하는 검사의 신뢰도 계수. 일반적으로 알파는 신뢰도의 하한값[92]이다.

---

[89] 이 지침서의 번역에서는 시험과 설문을 포괄하고자 "응답자"라는 용어를 선호하였다.

[90] 이 알파계수는 L. J. Cronbach의 1951년 Psychometrika 논문에서부터 널리 사용되기 시작했으나, 이미 1945년 L. A. Guttman의 Psychometrika 논문에서 λ_3 란 명칭으로 제시되었다. 따라서 Cronbach 본인도 자신의 이름으로 불리는 것에 부담스러워 했고 학계에서도 Guttman-Cronbach alpha 또는 alpha coefficient 로 부르고 있다.

[91] 검사이론에서는 '타우 동등검사'라고 한다.

[92] 따라서 알파계수를 신뢰도 추정치로 부르는 것은 문제가 있다. 단지 모든 Split-half 계수의 평균이 된다는 것 때문에 신뢰도 추정치로 오도되는 면이 있다.

역번역 설계[93]**.** 이 설계를 사용하면 한 번역자 집단이 검사의 원본언어판에서 대상언어판으로 변환한 다음 대상언어판을 두 번째 번역자 (또는 번역자 집단)가 원본언어로 역번역한다. 원본 검사와 역번역판을 비교하여 대상언어판의 적절성 여부를 판단한다. 원본과 역번역본이 매우 유사한 경우 대상언어판을 사용할 수 있다고 가정한다.

오메가. ("오메가 계수" 또는 "McDonald의 오메가"로 불린다) 하나의 속성을 공통으로 측정하는 문항들로 이루어진 검사(일반 요인 모형을 합치시킴)의 신뢰도 계수. 알파 계수보다 더 넓게 적용 가능하다[94].

원본언어. 원본 검사의 개발에서 사용된 언어.

차별기능문항(**DIF**; Differential Item Functioning). 어떤 문항이 두 개의 다른 집단에서 대부분 같은 기능을 하는지 결정할 수 있는 통계적 절차들이 있다. 먼저, 검사가 측정하는 속성의 잣대에서 응답자들을 대응시킨 후 해당 문항에 대한 수행을 집단 간에 비교한다. [그 수행(점수)에서] 집단 간 차이가 관찰되면, 그 문항이 편향될 가능성이 있다고 한다. 두 대응집단 간 점수차이는 조건화된 차이[95]인데 연구자는 이 차이를 설명하기 위해 노력을 해야 한다.

차별기능문항 식별을 위한 로지스틱 회귀분석. 이 통계적 절차는 DIF 분석을 수행하기 위한 한가지 방법이다. 집단별 검사 수행 자료에 로지스틱 곡선을 합치(fit)시키고 집단 간에 그 곡선들을 통계적으로 비교한다.

차별기능문항 식별을 위한 맨텔-헨젤 절차. 하나의 문항에 대한 두 집단의 수행을 비교하기 위한 통계적 절차이다. 집단 간 비교는 검사에서 측정되는 속성이나 구성개념의 차원에서 집단 간에 응답자들을 대응시켜 이루어진다.

쿠더-리처드슨 공식 20. (또는 간단하게 "KR-20"이라고 한다) 이분 문항으로 구성된 검사에서 문항들이 하나의 공통 속성을 측정하고 동일한 문항 변별도를 가진다는 가정하에 산출되는 신뢰도 계수이다[96].

---

[93] 순번역 결과에 대한 검토없이 역번역을 하는 것에 대한 문제가 제기될 수 있다.

[94] 알파계수에는 모든 문항의 변별도가 동일하다는(타우 동등검사) 제약이 있는 반면 오메가 계수에서는 문항들간 변별도가 다른 것을 허용하기 때문이다.

[95] 각 대등집단의 검사점수는 측정되는 "속성"의 잣대에서 "같은"위치에 대응되는 응답자들의 점수이므로 "속성"이 통제되고 있는 것이다. 따라서 그런 조건하에 얻어진 검사점수이기에 조건화된 점수이고 집단간 점수차이는 조건화된 차이가 된다.

[96] 이분문항점수에 대한 알파계수라고 할 수 있다.

탐색적 요인 분석. 탐색적 요인 분석은 검사(또는 [소검사들로 이루어진] 검사집)에 있는 문항들 간의 관계에 의해 생성된 상관 행렬을 이용한 통계 절차이다. 목표는 문항(또는 소검사) 간의 상호 연관성을 소수의 요인으로 설명하고자 하는 것이다. 예를 들어, 수학 시험에 대한 요인 분석은 문항들을 계산, 개념 그리고 문제-해결이라는 3개의 군집으로 분류하는 것이다. 즉, 요인분석을 통해 수학 시험은 계산, 개념, 문제-해결이라는 세 가지의 요인을 측정한다고 말할 수 있다.

현지화. 검사 분야에서 사용하는 용어로서, 하나의 언어와 문화에서 사용된 검사를 다른 언어와 문화에서 사용할 수 있도록 준비하는 과정을 기술하는 데 사용된다. 동등한 용어는 번역/번안이다.

확인적 요인 분석. 검사의 [내적] 구조에 대한 가설을 세운 다음, 검사에서 문항 간의 상관 [공분산] 행렬로부터 검사 구조를 얻기 위해 분석을 수행한다. 가설 구조와 추정된 구조가 충분히 밀접해서 두 구조가 같다는 귀무가설이 유지되는지 보기 위해 통계적 검정을 한다.