한국심리학회지: 일반
The Korean Journal of Psychology
2003. Vol. 22, No. 2, 57-80

심리검사 번안에 대한 통합적 접근

손 원 숙†
한국교육과정평가원

본 연구의 목적은 심리검사 번안 절차 시 연구자들이 고려해야 될 심리측정학적 사항을 종합적으로 검토해 보고자 하는 것이다. 이를 위해서 우선 국제검사위원회에서 발표한 「검사번안 지침서」를 소개하였고, 보다 타당하고 신뢰로운 검사 번안을 위해 사용할 수 있는 질적 및 양적인 절차들을 살펴보았다. 본 연구에서는 16PF 성격검사의 한글판과 영문판 검사간 점수 동등성(score equivalence)을 평가하기 위하여 두 가지 절차를 적용해 보았다. 먼저, 문항 수준에서의 동등성을 평가하기 위하여 차별기능문항기법 중 로지스틱 판별분석을 응용하였고, 차별기능문항의 원인을 탐색하기 위하여 문항평가질문지를 이용한 질적인 분석을 시도하였다. 마지막으로 우리 나라 실정에 적합한 심리검사 번안절차와 검사 사용에 대한 지침서 개발의 필요성을 논의하였다.

주요어: 검사번안, 점수 동등성, 차별기능문항기법

현재 우리 나라 뿐 아니라 많은 다른 국가에서도 다른 나라의 언어로 되어 있는 교육 및 심리검사를 자국의 언어로 번안 (adaptation)[1] 하여 사용하는 경향이 증가되고 있다(Hambleton, 1993;1994). 이러한 경향의 증가는 크게 세 가지 이유로 정리해 볼 수 가 있다(Hambleton & Kanjee, 1995). 첫 째, 비교문화 연구장면에서 서로 다른 언어를 사용하는 사람들의 지식, 기술이나 특성을 비교, 연구하려는 목적으로 검사를 번안하게 된다. 특히 21세기의 국제화 추세에 발맞추어 국가간 경제적, 교육적 혹은 문화적 교류에 적합한 검사 개발은 무엇보다도 중요한 흐름으로써 주목을 받고 있다. 최근에 여러 나라 학생들의 학력을 비교하려는 수학, 과학 성취도 추이변화연구(TIMSS-2003)나 OECD 학업성취도 국제비교

† 교신저자: 손원숙, 한국교육과정평가원, 서울 종로구 삼청동 25-1 별관404호, wsohn@kice.re.kr

1) 이 논문에서는 번역(translation)과 번안(adaptation)이라는 용어를 구분하여 사용할 것이다. 번안은 검사가 사용되는 문화적인 맥락을 고려하며, 번역의 과정을 포함하는 보다 포괄적인 용어이고, 한 언어에서 또 다른 언어로 검사를 바꾸는 과정을 보다 정확하게 설명한다고 볼 수 있다. 반면 번역은 타당한 검사를 만들기 위한 여러 번안 과정 중 한 단계로 볼 수 있다.

연구 (PISA-2003)등이 대표적인 예가 될 것이다. 또한 비교심리학자들의 기본적인 관심사인 특정한 구인의 보편성 혹은 특수성을 검토하기 위하여 이러한 번안절차는 이뤄진다. Binet-Simon의 지능검사는 이미 1911년에 불어에서 영어로 번안되었고, 미네소타 다면성격검사와 같은 많은 성격검사들 역시 다양한 언어판 검사들이 존재한다(Butcher, 1996). 두 번째로, 심리학 연구 일반에서 어떤 하나의 구인(construct)을 측정하고자 하지만 자국에서 개발된 검사가 없는 경우, 시간·경제적인 이유로 외국의 검사들을 번안해서 사용하기도 한다. 검사 개발 자원이나 경험이 부족한 상황에서 이미 개발되어 널리 사용되고 있는 다른 나라의 검사들을 번안하게 된다. 마지막으로 검사의 공정성(fairness)을 높이기 위한 목적으로 검사는 번안되어 사용된다. 예컨데, 자격증시험이나 회사의 승진시험과 같은 고부담 검사(high-stakes testing)에서 피평가자들이 다국어, 다민족 집단인 경우, 검사결과의 공정성을 위하여 검사는 반드시 다국어로 번안되어 사용되야 할 것이다. 특히 많은 전문인력을 필요로 하는 정보통신분야에서 이러한 경향은 두드러지고 있는데, 소프트웨어 회사인 NOVELL에서는 국적에 상관없이 능숙한 전문인력을 선발하기 위하여 12가지 언어로 번안된 국제자격시험을 운영하고 있다.

우리 나라의 대다수 연구자들은 새로운 검사를 개발하는데 소요되는 시간적·경제적 비용을 최소한으로 줄이기 위하여 외국에서 이미 검증되어 널리 사용되고 있는 외국검사들을 차용하고 있는 경우가 많다. 그러나 이러한 관행에서 연구자들이 간과해서는 안될 점은 한 언어에서 이미 검증된 검사라고 할지라도 그 검사가 한국어로 번역되어 졌을 때, 동일한 수준의 신뢰도나 타당도를 보장하지는 않는다는 점이다. 즉, 한 언어로부터 또 다른 언어로 검사를 단순히 번역하는 과정이 심리측정학적으로 동등한 검사를 산출하지 않는다는 점이다(Allalouf, Hambleton, & Sireci, 1999). 따라서 原검사와 번역된 검사 문항들이 동일한 의미를 가지며 동일한 문항특성을 지니고 있는지 등에 대한 동등성의 문제는 반드시 경험적으로 평가되어야 한다. 여기서 "동등성(equivalence)"이라는 것은 검사점수들이 서로 다른 집단에서 비교될 수 있는 측정수준을 말하는 것으로, 집단간 타당한 점수비교의 전제조건이라고 할 수 있다. Drasgow (1984)는 이런 동등성을 문항반응이론의 맥락에서 "관찰 점수와 그 검사가 측정하려고 하는 잠재속성간의 관계가 하위 집단에서 동일할 때 동등한 측정(equivalent measurement)이 얻어지는 것"이라고 정의 내렸다. 이러한 동등성을 확립하기 위하여 현재 여러 가지 경험적인 방법들이 사용되고 있는데, 특히 국제검사위원회 (International Test Commissions: ITC)에서 개발한 검사번안지침서에 따르면, 적절한 번역기법, 즉 판단적인 방법과 더불어서 통계적 방법을 사용하는 종합적인 접근법을 제안하고 있다 (Hambleton, 1993; 1994 참고).

본 논문에서는 우선 1995년 공식적으로 발표된 국제검사위원회의 검사번안지침서를 간략히 소개하면서, 연구자들이 심리검사 번안시 고려해야 될 구체적인 사항들에 대해서 논의할 것이다. 또한 검사간 동등성 확립을 위하여 사용되고 있는 판단적 및 통계적 절차를 소개할 것이다. 마지막으로 심리검사번

안을 위한 종합적인 접근이 실제자료에 어떻게 적용될 수 있는지 예시하고자 한다.

심리검사 번안 시 고려사항[2)]

검사 개발, 신뢰도 및 타당도 평가, 규준작성 등에 대한 기술적인 규준이나 지침서는 많은 국가들에서 연구되어졌으며, 대표적으로 AERA, APA, NCME의 *Standards for educational and psychological testing* (AERA, APA, NCME, 1985; 1999)은 각 나라의 언어로 번역되어 활발하게 적용되고 있다. 하지만 연구자들이 점수 동등성(score equivalence)을 확립하면서 외국 검사를 번안하고자 할 때, 참고할 수 있는 지침서는 최근까지 개발되지 않고 있었다. 이런 가운데 검사번안절차를 위한 타당한 지침서 개발에 뜻을 함께 한 13명의 심리학자들로 구성된 국제검사위원회는 1992년부터 2-3년간 연구한 끝에 하나의 지침서를 발표하게 되었다(Hambleton, 1994; Van de Vijver & Hambleton, 1996). 이 지침서는 크게 4가지 영역으로 구성된 22가지 지침을 포함하고 있다. 첫 번째 영역은 문화적인 맥락과 관련된 지침들로 구성되어 있고, 두 번째 영역은 검사 개발과 번안의 기술적인 측면들에 대한 내용을 포함하고 있다. 세 번째 영역은 검사 실시와 관련되고, 마지막 영역은 결과를 문서화하고 해석하는 일에 대한 내용을 포함하고 있다. 또한 최종 지침서에는 각 지침에 대한 설명, 각 지침을 따르기 위한 절차, 자주 범할 수 있는 오류들, 그리고 참고문헌들을 상세하게 담고 있다. 이 장에서는 국제검사위원회에서 개발된 지침들을 소개하면서, 연구자들이 심리검사를 번안할 때 고려해야 할 사항들에 대해서 논의하고자 한다.

맥락 (Context)

이 맥락 영역에 속하는 지침들은 原검사와 번안된 검사들 간 구인 동등성(construct equivalence)에 관한 문제를 다루고 있다. 구인 동등성이란 각 집단에서 동일한 구인이 측정되어진 경우로서, 문화-독립적인(culture-independent) 혹은 보편적인(universal) 구인 타당도를 의미한다 (Sireci, Bastari, & Allalouf, 1998). 반면, 구인 비동등성 (inequivalence)이란 그 검사가 여러 문화/언어 집단에서 서로 다른 구인들을 측정하고 있거나, 그 구인의 개념이 집단간에 공유되는 부분이 작은 경우일 것이다. 이 영역의 지침들에서는 집단 간 검사 동기나 검사에 대한 익숙함의 정도는 검사의 목적과는 직접적으로 관련이 없지만, 이러한 면에서 나타나는 문화적 차이는 검사의 수행에 큰 영향을 줄 수 있음을 지적하였다. 따라서 가능하면 이러한 문화적 차이에 대한 영향을 최소화 시킬 수 있도록 두 집단의 검사 환경을 동등하게 구성하는 것이 중요한 문제이다. 原검사가 번역될 때, 그 번역된 언어를 대상 언어(target language)라고 부르고, 대상언어판 검사를 사용하게 될 집단을 대상 집단(target group)이라고 한다. 검사 번안자들은 그 검사도구가 측정하고 있는 구인이 대상 집단(target group)에서도 동일한 형태로 인식되고 있는지 확인해 보아야 한다. 또한 그 구인에

2) Draft of the instrument adaption guidelines provided by International Test Commissions (ITC) (Hambleton, 1994 참고)를 참고함.

대한 정의가 집단들 간에서 어느 정도 일치하고 있는지를 고려해야 한다 (van de Vijver & Tanzer, 1997). 사실 구인 동등성이 이뤄지지 않은 상황에서 검사를 번안한다는 것 자체가 무의미하기 때문에 검사 번안 시, 구인 동등성에 대한 검토는 일차적으로 이루어져야 한다(Sohn, 2002).

검사도구 개발 및 번안(Instrument Development and Adaptation)

이 영역에 속한 10개의 지침들은 번역자를 선정하는 문제에서부터 실제 자료를 가지고 점수 동등성을 평가하는 통계적인 기법에 이르기까지 검사를 번안하는 전체 과정에 적용될 수 있는 내용을 포함하고 있다.

번안의 문제

우선 原검사를 번안하는 과정에서는 原검사 문항을 그대로 직역하기보다는 대상 집단의 문화적 차이를 충분히 고려하는 방식으로 검사 번안이 이루어져야 한다. 이를 위해서는 적절한 번역자의 선정 문제는 상당히 중요하다. 예를 들어서 미국의 지능검사를 한국어로 번안한다고 하자. 이 경우 검사 번역자로서의 기준은 영어와 한국어 모두에 능숙해야 함은 물론이고, 각 나라의 문화와 검사개발 원리 및 문항 제작에도 지식이 있어야 한다. 더불어서, 영역 지식(domain knowledge) 즉, 지능 및 심리학에 대한 기본적인 지식 역시 필요하다. 실제로 해당 영역과 검사개발에 대한 지식이 없는 사람이 심리검사를 번역하는 경우는 그 번역이 오로지 문자적인(literal) 번역에만 그쳐서, 그 번안된 검사가 피검자들에게 오해를 불러 일으킬 수 도 있다. 이처럼 검사를 번안하는 일은 단순한 절차가 아니기 때문에 가능하다면 여러 전문가들로 구성된 팀 접근법(team approach)이 유용할 것이다. 이 번역의 절차에 대해서는 다음 장에서 판단적 방법을 소개하면서 좀 더 자세히 논의하기로 하겠다. 또한 검사에서 사용되는 단어들은 난이도 수준, 읽기 쉬운 정도(readability), 문법, 쓰기 양식 및 구두법 면에서 집단들 간에 동등해야 하며, 검사 기법, 문항 형태 및 내용, 검사 절차들이 대상집단에서도 原집단에서와 마찬가지로 친숙해야 한다. 경우에 따라서 사지선다형이나 서답형 문항과 같은 특정 문항 형태, 도표 제시 방식이나 지필 혹은 컴퓨터화 검사 등과 같은 요인들이 모든 검사대상자들에게 동일한 정도로 익숙하지 않을 수도 있다는 점을 주의해야 한다. 또한 이 영역에서는 原검사와 번안된 검사간의 동등성을 확립하기 위한 판단적 방법의 사용을 강조하였다. 비교문화연구에서 사용되는 대표적인 번안기법들로는 선번안(forward-adaptation) 기법과 역번안(back-adaptation) 기법이 있다 (Hambleton, 1993;1994) (※이 기법들에 대한 설명은 다음 장에서 구체적으로 논의할 것이다). 실제적으로 이 판단적 방법들은 검사가 시행된 후 통계적인 기법이 적용되기 전에 번역의 동등성을 검토하는 선행적인 절차로서 이용되고 있다.

자료수집설계

ITC 지침서에서는 판단적인 방법과 더불어서 검사간 동등성을 검토하기 위하여 실제 자료를 모으고, 그에 대한 통계적 분석을 실시하는 절차를 소개하고 있다. 우선 통계기법

들을 적절하게 적용할 수 있기 위해서는 자료수집절차를 설계해야 한다. 자료수집과정은 우선 의미있는 해석이 가능하도록 충분히 큰 표본을 확보해야 한다. 특히 여러 언어판 검사들 간 동등성 검토를 위하여 사용되는 대부분의 통계적 기법들(예: 구조방정식 모형, 문항반응이론, 혹은 문항편파 연구들)은 충분히 큰 표본을 필요로 한다. 그러나 많은 경우 대상 언어 표본은 原언어 표본에 비하여 그 크기가 작아서 신뢰롭고 타당한 통계적 분석이 어려워지기도 한다. 여러 종류의 언어판 검사간 동등성을 확립하기 위하여 사용될 수 있는 자료 수집 설계는 크게 세 가지로 요약할 수 있다(Hambleton, 1993, 1994; Sireci, 1997): (a) 단일언어구사집단 설계 (separate monolingual group design), (b) 이중언어구사집단 설계(bilingual group design), (c) 결합 단일언어구사집단 설계 (matched monolingual group design). 이 세 설계는 포함된 피험자의 유형(단일언어 구사자 혹은 이중언어 구사자), 실시되는 검사유형(原검사, 번역된 검사, 혹은 역번역된 검사), 그리고 사용되는 특정한 통계기법에 따라서 분류되었다. 가장 광범위하게 사용되는 첫 번째 설계는 原언어 집단 구성원들에게 原검사를 실시하고, 대상언어 집단 구성원에게 대상언어판 검사를 실시하도록 하는 것이다. 이 설계의 장점은 결과의 일반화가 용이하다는 점이다. 그러나 서로 다른 집단이 사용되기 때문에, 검사 점수가 표본들간 피검사자들의 특성 차이와 혼돈될 수 있다는 단점이 있다. 두 번째 설계는 原언어와 대상언어 모두에 능숙한 이중 언어 구사자들에게 原검사와 번안된 검사 모두를 실시하는 것이다. 이 설계에서는 첫 번째 설계와는 달리 동일한 피험자 집단을 대상으로 하기 때문에, 피검사자들의 특성과 검사점수와의 혼돈현상을 통제할 수 있다는 장점이 있다. 기본적으로 이 설계에서는 이중언어 구사자들이 양언어에 동일한 정도로 능숙하다는 가정을 갖는다. 하지만 현실적으로 많은 이중 언어 구사자들이 양 언어에 대해서 동일한 정도의 능숙함을 갖기는 힘들다. 또한 결과 일반화의 문제로서 이중 언어 구사자들을 대상으로 한 결과를 단일 언어 구사자들에게 적용시킬 수 있는지에 대한 한계이다. 마지막 설계에서는 原언어 집단 구성원들에게 原검사와 역번역된 검사를 시행한다. 이 설계의 장점은 두 번째 설계와 마찬가지로 피검사자들의 특성과 검사점수와의 혼돈현상을 통제할 수 있다는 점이다. 반면 이 설계에서는 기본적인 목적이 검사 점수의 의미를 대상 언어집단 구성원들에게 일반화시키기 위함인데 번안된 검사에 대한 자료를 얻을 수 없다는 제한점이 있다. 또한 연습효과에 의해서 두 검사의 수행이 독립적이지 않을 수 있지만, 이 경우 상쇄 균형화(counterbalancing)로 어느 정도 연습효과를 배제할 수 있을 것이다.

통계적 기법의 적용

자료를 수집한 후, 연구자들은 검사들간의 동등성을 확립하려는 목적으로 적절한 통계기법을 사용하게 된다. 우선 통계적 기법들은 검사들간의 동등성을 평가하기 위해 유용한 정보를 제공해 줄 수 있는데, 보통 판단적 방법을 보충해 줄 수 있다는 점에서 두 기법은 상호보완적으로 사용된다. 또한 통계적 기법은 실제 검사가 실시되는 상황에서 피검사들로부터 직접적인 정보를 얻을 수 있다는 장

점을 지니고 있다. 이러한 동등성 평가에 사용될 수 있는 통계적 기법은 크게 두 가지로 분류될 수 있다. 우선 검사들간에 공통척도(common scale)가 가정된 경우로서, 검사간 점수가 동일한 의미를 갖는다고 생각하고 직접적인 검사간 점수비교를 하는 것이다. 주로 요인 분석적 방법들이 대표적이다. 또한 조건적인 기법(conditional procedure)들이 사용되는 경우로서, 피험자들을 특성 수준별로 분류하고, 동등성은 각 수준에 따라서 평가되어진다. 다음 장에서 소개할 차별기능문항기법들이 조건적인 기법이라고 볼 수 있다.

이러한 과정들을 통해 대체적으로 잘못 번안된 문항들이나 문화적으로 적절하지 않은 문항들이 두 검사간에 동등하지 않은 문항들로 파악될 것이다. 이런 문항들은 집단간 비교를 할 때는 사용할 수 없으며, 잘못 번안된 문항들의 경우는 수정되거나 제거될 수 있다. 그러나 번역상으로는 문제가 없지만 동등하지 않은 것으로 판명 난 문항들은 특정한 집단에 대한 독특한 정보를 제공해 줄 수 있을 것이다. 따라서 각 문항이 왜 동등하지 않은지에 대한 원인을 연구함으로써 각 문화/언어 집단에 대한 추가적인 통찰을 얻을 수 있을 것이다.

검사 실시 (Administration)

세 번째 영역에서는 검사를 실시하는 절차와 관련된 모든 사항을 포함하고 있는데, 검사 실시자 선택에서부터 문항형태와 검사 실시시간 선정 등에 관한 내용과 관련된다. 먼저 검사 실시자는 점수로부터 도출된 추론의 타당성을 약화시킬 수 있는 제시자료, 실시절차, 그리고 반응 양태와 관련된 모든 요소들에 대해서 민감하게 반응해야 한다. 가능하면 빈번하게 나타날 수 있는 문제들이나 타당도를 위협하는 요인들에 대해서 목록 등을 작성해 놓는 것이 바람직하다. 검사 실시자들에 대한 철저한 훈련 및 상세한 실시 요강 등을 마련하는 것 역시 꼭 필요한 절차이다 (Poortinga & van de Vijver, 1987). 실시자의 성별, 나이 혹은 의상 스타일 역시 영향을 끼칠 수 있으므로, 대상 집단에 속한 정보원들의 도움을 받거나 예비검사 등을 실시하여 이에 관한 피험자들의 정보를 수집하는 것이 좋다.

문서화/점수해석(Documentation/Score Interpretations)

마지막 영역에서는 번안과정에 대한 문서화와 점수 해석에 대한 내용을 포함하고 있는데, 이 과정은 흔히 등한시 할 수 있으나 반드시 필요한 절차이다. 번안된 검사의 경우, 그 번안 절차와 原검사와의 동등성 검토 결과를 검사요강에 문서화해야 한다. 예컨대, 사용된 자료수집설계, 번안 및 동등성 평가방법, 번역가의 선정 및 훈련, 특정한 문항을 포함 혹은 삭제한 경우 그에 대한 이유, 번안과정에서 나타난 문제들, 이에 대한 해결방법, 검사 실시와 관련된 사항 등 모든 절차들을 상세히 기록해야 한다. 번안된 검사를 이용하여 대상집단과 原집단간의 점수비교를 할 때에는 두 언어판 검사간에 확립된 동등성의 정도에 근거하여야 한다. 경우에 따라서 여러 언어판 검사간에 공통 척도를 형성하여 집단간 모든 수준의 점수비교가 가능할 수

도 있다. 하지만 여러 언어판 검사들간에 적절한 동등화 과정이 적용되지 못한 경우 점수들간의 직접적인 비교는 타당하지 않음을 유의해야 한다.

동등성 평가 방법

판단적 방법 (Judgmental Methods)

비교문화 연구자들이 제안한 검사 번안기법들로는 탈중심화기법(decentering), 위원회 접근법(committee approach), 역번안기법, 선번안기법 등이 있다(Brislin, 1970). 이 중에서 심리학이나 교육학 일반에서는 역번안기법과 선번안기법이 가장 광범위하게 사용되고 있으며, 여러 가지 경험적인 연구에 의하면 역번안기법이 번역초기과정에서 번역의 질을 평가하는 데 효과적이라고 알려져 있다 (Hulin, Drasgow, & Komocar, 1982; 김아영 · 임은영, 2003). 역번안기법은 먼저 검사 도구가 대상 언어로 번역되고, 그런 후 다른 번역자(들)에 의해 그 번역된 검사는 다시 原검사 언어로 전환된다. 최종적으로 原검사와 역번안된 검사 문항을 서로 비교하여 동등성이 평가된다. 이러한 절차는 만족할 만한 수준의 번안된 검사가 나올 때까지 지속된다. 이 역번안기법은 동등성 평가를 위한 검증된 효과에도 불구하고 몇 가지 단점을 지니고 있다. 먼저 번역의 질을 평가하기 위하여 이중 언어 구사자를 이용한다는 점인데, 현실적으로 두 언어 및 문화에 동일하게 능숙한 이중 언어 구사자 선발이 쉽지 않다는 점이다. 또한 이러한 이중 언어 구사자들이 단일 언어 구사자들과 동일한 방식으로 검사 문항들에 반응할 것인지에 대해서도 고려해 보아야 한다는 점이다. 또한 역번역자들이 비록 1차 번역본이 잘못 만들어졌더라도 그들은 이미 양쪽 언어에 익숙하므로, 역번역본을 만들 때, 그것을 올바르게 고쳐서 역번안 할 가능성도 배제할 수 없다. 마지막으로 동등성에 대한 평가가 原언어에 대해서만 이루어지기 때문에 이 결과를 대상언어판 검사에 일반화시키는 데에는 한계가 있을 수 있다는 단점이 있다. 이러한 단점에도 불구하고, 많은 연구자들은 역번안기법은 연속적인 번안-역번안 과정에 의하여 原검사와 가장 유사한 번안된 검사를 만드는데 효과적이라고 밝히고 있다(Werner & Campbell, 1970).

역번안기법과 더불어서 많이 사용되고 있는 선번안기법은 먼저 그 검사를 번역한 후, 검사간 동등성은 또 다른 번역자들에 의하여 판단 · 수정되는 방식이다. 이 방법의 장점은 번안된 검사에 대한 이해도나 수용가능성에 관한 정보를 직접적으로 얻을 수 있다는 점이다. 반면, 이 검사 번안 절차는 검사 동등성을 평가하기 위하여 번역자들이 原검사 문항의 의미를 이해해야만 번안된 문항에 대한 판단이 가능하다는 단점이 있다.

특별히 성격검사와 같은 심리검사를 번역하는 문제는 심리적이고, 언어적이며, 문화적 고려가 모두 필요한 아주 어려운 과정이다. 실제로 검사를 번안하려고 할 때는 Casagrande (1954)가 제안한 실용적, 심미적, 문화인류학적, 언어적 번역의 목적이 모두 고려되어야 한다. 이는 교육 및 심리검사를 번안하여 사용하는 것은 단순히 언어적인 번역만을 제공하는 것 이외에 그 검사가 재고자 하는 구인에 대한 이해를 바탕으로 하는 번안 과정이

필요하다는 사실을 시사하고 있다(Hulin, Drasgow, & Parson, 1983).

통계적 방법 (Statistical Methods)

최근까지 많은 비교문화연구자들은 原검사와 번안된 검사간 동등성을 평가하기 위하여 역번안기법과 같은 판단적인 방법을 주로 사용하였다. 이 역번안 기법은 여러 개의 경험적인 연구들에서 밝혀졌듯이 번역의 질을 평가하는 초기작업에는 상당히 유용한 것으로 나타났지만 (Hulin, et. al, 1982), 비교 문화 및 언어의 동등성을 수립하는 데에는 충분하지 않으며, 이와 더불어서 통계적인 방법의 필요성이 제기되어 지고 있다. ITC 지침에서도 "검사 개발자들은 여러 언어판 검사의 동등성을 평가하기 위하여 적절한 통계적인 방법을 질적인 방법과 더불어서 사용하기를 권장한다"고 하였다. 또한 국내의 한 연구(김아영 · 임은영, 2003)에서도 검사의 동등성을 확립하기 위한 세 가지의 방법(단순번역과 검토, 역번역과 검토, 역번역, 검토와 경험적인 타당도검증)을 비교해 본 결과, 역번역기법과 경험적인 타당도 연구를 함께 사용한 경우에 가장 바람직한 결과를 산출하였다고 보고하였다.

현재 原검사와 번안된 검사간의 동등성을 문항 수준에서 평가하기 위하여 사용할 수 있는 통계적인 방법으로 차별기능문항(이하 DIF라고 부름) 기법이 소개되고 있다 (Drasgow & Probst, 2000). 이런 연구를 통하여 연구자들은 번역의 질과 서로 다른 집단간 검사 점수의 동등성에 대한 정보를 얻을 수 있다. 또한 문제를 나타내고 있는 문항들을 파악할 수 있다는 점에서 DIF 기법들은 상당히 유용하게 응용되고 있다. 서로 다른 언어 혹은 문화 집단에 속하는 개인들이 검사가 재고 있는 속성, 예컨데 능력, 성격, 혹은 태도등이 동일하다면, 이 두 개인들은 같은 문항에 대하여 동일한 방식으로 응답할 것이다. 그러나, 만약 그들이 동일한 방식으로 응답하지 않는다면 그 문항은 "DIF를 나타낸다"고 할 수 있으며, 그 원인에 대한 심층적인 분석이 필요할 것이다.

이처럼 검사 문항간 동등성을 평가하기 위하여 최근에 다양한 DIF 기법들이 사용되고 있는데, 문항반응이론에 근거한 차별 문항 및 검사 기능 접근법 (DFIT)을 이용한 연구(Ellis & Mead, 1998; Hulin et. al, 1982)도 있으며, SIBTEST(Gierl & Khaliq, in press), 멘텔-헨젤 기법 (Allalouf, et. al, 1999), 또는 로지스틱 회귀분석(Gierl, Rogers, & Klinger, 1999)등과 같은 DIF 기법들도 광범위하게 응용되고 있다. 이 중 본 연구에서는 Miller와 Spray(1993)가 로지스틱 회귀분석의 하나의 변형으로서 제안한 로지스틱 판별분석(이하 LDFA로 부름)을 소개하고자 한다.

로지스틱 판별분석 (Logistic Discriminant Function Analysis)

LDFA(Miller & Spray, 1993)의 판별함수는 다음과 같다.

$$\text{Prob}(G \mid X, U) = \frac{e^{(1-G)(-\alpha_0-\alpha_1 X-\alpha_2 U-\alpha_3 X*U)}}{1+e^{(-\alpha_0-\alpha_1 X-\alpha_2 U-\alpha_3 X*U)}} \quad (1)$$

여기에서 G는 소속 집단, X는 대응변수(matching

variable)의 점수, U는 각 문항의 응답, 그리고 X*U는 X와 U, 두 개 변인의 積이다. 이 판별함수의 계수는 α(i=0, 1, 2, 3)로서 나타내고, 이들은 우도 함수를 최대화시키는 방식으로 추정된다.

대응변수(X)는 DIF를 추정하기 앞서서, 집단간 비교 가능성을 보장하기 위해 사용되는 대응을 위한 준거 변수로서, LDFA에서는 검사 총점을 사용한다. 공식(1)에서 소속 집단, G=1 이면 참조집단, G=0이면 초점집단을 의미한다. DIF 분석 시 비교하게 되는 두 피험자 집단 중 일종의 기준이 되는 집단을 참조(reference) 집단이라고 하고, 반면 연구의 관심 대상이 되는 집단을 초점(focal) 집단이라고 부른다. 참조집단의 검사 결과를 기준으로 초점집단에서 차별적인 기능을 하는 문항을 추출하게된다. 통상적으로 남녀 혹은 인종간 DIF를 분석하게 되면, 여자 집단 혹은 흑인 등 소수민족 집단 등이 초점집단이 되는 경향이 있다. 본 연구에서는 한국집단이 초점집단으로, 미국집단이 참조집단으로 정의되었다.

공식 (1)에서 각 문항의 반응, U는 두 개 이상의 범주를 모두 포함시킬 수 있어서 이분 문항 뿐 아니라, 다분 문항도 이 함수에서 다룰 수 있다. DIF는 방향성에 의하여 두 가지 종류로 분류될 수 있는데, 일방적 혹은 균일적 (uniform) DIF와 비일방적 혹은 비균일적(nonuniform) DIF이다. LDFA에서는 이 두 가지 종류의 DIF를 모두 다룰 수 있다. 일방적 혹은 균일적 DIF라는 것은 한 집단의 피험자들이 또 다른 집단의 피험자들 보다 정답 혹은 긍정 반응 확률이 모든 잠재특성 범위에서 균등하게 높게 (혹은 낮게) 나타내는 경우를 말한다. 반면 비일방적 혹은 비균일적 DIF는 집단과 잠재특성간에 상호작용이 존재하는 경우로서, 잠재특성의 범위에서 두 집단의 정답반응(혹은 긍정) 확률의 차가 일정하지 않은 경우이다. 판별 계수 α_2와 α_3의 유의성에 대한 우도비검정은 일방적 DIF와 비일방적 DIF를 평가하는 것으로, 만약 $\alpha_2 \neq 0$ 고, $\alpha_3 = 0$ 이면, 그 문항은 일방적 DIF를 나타내는 것이고 $\alpha_2 = 0$ 이고, $\alpha_3 \neq 0$ 이면, 그 문항은 비일방적 DIF를 나타내는 것이다.

구체적으로, 이 LDFA에서는 세 가지의 모델이 검증되어지는데 첫 번째, 완전모형은 대응점수, 문항, 그리고 대응점수와 문항의 상호작용으로 구성되어 있고 (공식 1 참고), 두 번째, 축소 모형은 완전모형에서 상호작용만이 제외되고, 마지막으로 영모형은 축소모형에서 문항 점수가 제외된, 즉 대응변수만 고려한다. 따라서, 일방적인 DIF을 검사하기 위해서 축소모형과 영모형의 우도함수를 비교하고, 비일방적인 DIF를 평가하기 위해서는 축소모형과 완전모형의 우도함수를 고려한다. 일단 일방적 혹은 비일방적 DIF가 유의도 검증에서 밝혀졌다면, 적어도 문항 반응의 한 수준(16PF의 경우는, U=0, 1, or 2)에서 어떤 집단의 확률은 문항점수와 대응변수 점수가 주어졌을 때, 오로지 대응변수 점수만으로 예측했던 것과는 유의하게 달라진다라는 결론을 내릴 수 있다.

이런 유의도 검증을 실시한 후, 통계적 유의성과는 별도로 DIF의 실제적 심각성(practical severity)과 DIF의 방향을 알아보기 위하여 그래프 기법을 활용한다. 즉, 유의한 DIF를 보였던 문항들을 대상으로 각 문항 점수 수준 (U=0, 1, 2)에서 추정된 판별함수 주변에 95%

Scheffe 유형의 신뢰구간을 형성하고, 영모델 판별함수와 비교하는 것이다. 영모델이라는 것은 문항 반응이 제외된 모델로서, 이는 어떤 문항에서든 동일하여, 이는 no-DIF 회귀(regression)로서의 역할을 한다. 만약 이 신뢰구간이 대부분의 대응변수 점수에서 영모델을 포함한다면, 그 문항이 통계적으로 유의한 DIF를 보였더라도, 그 DIF는 실제적으로 심각한 정도는 아니라고 간주한다.

Potenza와 Dorans (1995)는 차별기능문항 기법들을 크게 두 가지 차원에 의하여 분류하였다. 첫째로, 대응변수로 사용되는 특성(trait) 추정치의 성격, 즉 그것이 관찰된 점수인지 잠재 변인인지에 따라서 관찰 점수 접근법과 잠재 변수 접근법으로 분류하였다. 두 번째로는 각 특성의 수준에서 문항 수행이 결정되는 방식이 수학적 함수에 의해 추정하는 것이면, 모수적 접근, 그렇지 않으면 비모수적 접근으로 분류하였다. 이 분류에 따르면, LDFA는 관찰된 점수를 대응변수로 사용하는 모수적 기법으로 분류된다. 다른 DIF 기법들과는 달리, LDFA는 처음부터 다분문항의 사용을 위하여 개발되어 진 것으로, 표본의 크기만 적절히 크다면, 다분문항에서 DIF를 추출하는데 상당한 검증력을 지니고 있다. 또한 이 기법은 두 가지 유형의 DIF 모두를 추출할 수 있고, 한 개 이상의 대응변수를 다룰 수 있는 신축성이 있으며, 시각적으로 DIF의 정도와 방향을 알 수 있도록 해주는 그래프 기법을 제공한다는 장점이 있다. 그러나, 이 기법은 적절하게 모수를 추정하기 위하여서 상당히 큰 표본을 필요로 한다는 단점이 있다(i.e., n > 1500). 본 연구는 비교적 큰 표본을 가지고 있고, 다분 문항으로 구성된 검사 문항을 사용한다는 점에서 LDFA의 사용이 적절하다고 판단된다.

연구방법

자료수집설계 및 연구대상

이 연구에서는 단일언어집단설계(Separate monolingual group design)에 의하여 문화와 언어가 다른 두 개의 집단, 한국 대학생 538명(약 60% 여자, 약 40% 남자)과 미국대학생 844명 (약 70% 여자, 약 30% 남자)을 피험자로 사용하였다. 미국 대학생들에게는 原검사를 실시하고, 한국대학생들에게는 번안된 한글판 16PF 검사를 실시하였다. 본 연구에서는 한국과 미국집단에서 비교 가능한 표본을 선정하기 위한 목적으로 가능한 한 같은 직업군에 속하고, 비교적 비슷한 교육 수준을 가지고 있다고 판단되는 대학생 집단을 피험자로 선정하였다.

측정 도구

이 연구에서는 정상 성인을 위한 성격검사로서 가장 광범위하게 쓰여지는 검사 중 하나인 The Sixteen Personality Factor (16PF) Questionnaire (Cattell & Cattell, 1995)가 사용되었다. 이 16PF 검사는 한국에서는 염태호와 김정규(1990)에 의하여 표준화되어서 "성격요인검사"라는 이름으로 사용되고 있지만, 이 검사는 영어판 16PF와 비교해 볼 때 문항의 수, 문항반응 양식, 및 내용 면에서 상당 부분 수정되어 있다. 따라서, 문항 수준에서의 동등성을 검토하고자 하는 본 연구의 목적에

는 적절치 않았기 때문에, 본 연구에서는 Shaughnessy와 Kang(1998)이 영재연구를 위하여 사용하였던 한글판 16PF 검사를 수정하여 사용하였다. 16PF 검사는 세 가지 수준의 강제 선택 문항 양식을 사용하는데, 즉, "그렇다", "잘 모르겠다 혹은 둘 다 그렇지 않다", 그리고 "그렇지 않다"라는 방식의 반응양식을 가지고 있다(단 추리척도는 제외). 이 16개의 주요요인들은 외향성, 불안감, 강인함, 독립성과 자기통제 등 5가지의 이차 요인들로 다시 묶여질 수가 있다. 이 이차 요인들 가운데, 본 연구에서는 51개의 문항으로 구성된 외향성 요인만을 살펴볼 것이다. 외향성 요인은 다섯 가지 하위척도로 구성되어 있는데, 온정성 (factor A: 11문항), 쾌활성 (factor F: 10문항), 사회적 대담성 (factor H:10문항), 내밀성 (factor N:10문항), 자기 의존성 (factor Q2:10문항)이다.

연구 분석 절차

번역절차

Shaughnessy와 Kang (1998)이 마련한 한글판 16PF 검사를 기초로 하여, 일종의 선번안 기법을 응용하여 최종 한글판 16PF 검사를 마련하였다. 번안절차는 우선 검사 번역자로서의 기준에 부합되는 두 명의 박사과정 학생들을 선정하였다. 검사 번역자로서의 기준은 영어와 한국어 모두에 능숙해야 함은 물론이고, 두 나라의 문화와 검사개발 원리 및 문항 제작에도 지식이 있어야 한다. 또한, 성격 및 심리학에 대한 영역(domain)적 지식을 가지고 있어야한다. 이들은 Shaughnessy와 Kang의 한글판 16PF 검사의 번역 수준을 평가하고, 두 언어판 검사에서 동등하지 않은 문항들이 발견되면 이를 수정하도록 하였다. 여러 번의 편집 회의를 거쳐서, 이 두 명의 번역가로부터 최종 한글판 16PF 검사 문항들이 얻어 졌다.

문항의 채점

16PF 검사 문항들은 검사 매뉴얼(Russell & Karol, 1994)에 근거하여 채점이 이루어졌으며, 정답으로 정해진 반응들은 "2점"을, 중립반응은 "1점" 그리고 나머지 반응들은 "0점"이 주어졌다.

일차원성 (unidimensionality) 평가

차별기능문항 분석에 앞서서 자료의 차원성(dimensionality)이 평가되었다. 만약 문항들이 그 검사가 재고자 하는 구인 이외의 다른 것을 재고 있다고 한다면 그것은 문항편파를 일으키는 하나의 원인이 될 수 있기 때문에 일차적으로 차원성을 평가하였다. 본 연구에서는 인종 집단 별로 각 척도의 차원성을 평가하기 위해서 주성분 분석과 다분문항을 다룰 수 있는 Poly-DIMTEST(Stout, 1987)를 사용하였다. 즉, 검사를 구성하고 있는 문항들이 그 검사가 측정하고자 하는 능력, 성격 혹은 태도와 같은 하나의 구인(construct)을 재고 있는지 평가해 보았다. 각 인종집단별로, 외향성의 다섯 가지의 하위척도들은 개별적으로 평가되었으며, 먼저 PRELIS 2.0 (Jöreskog & Sörbom, 1986)을 이용하여 다분상관계수(polychoric correlations)를 바탕으로 주성분 분석이 실시되었다. DIMTEST는 검사 문항의 본질적인 일차원성(essential unidimensionality)을 통계적으로 평가하기 위한 프로그램으로

특정한 문항군이 다른 문항들에 비해서 본질적인 일차원성을 갖고 있다는 영가설 H_o: $d_e=1$과 다차원성에 대한 대안가설 H_a: $d_e>1$을 검증하게 된다(Stout, 1987).

차별 기능 문항 분석

통계 프로그램 SAS 8.1의 PROC LOGISTIC 절차를 DESCENDING 옵션과 함께 사용하여 LDFA를 실시하였다. 참조집단과 초점집단이 각각 "1" 그리고 "0"으로 코딩되어졌기 때문에, 판별계수가 양수이면 미국집단이 유리하게 그리고 음수이면 한국집단이 유리하게 기능하는 문항임을 의미한다. 유의한 DIF를 보이는 문항에 대해서는 판별함수 주변에 95% 신뢰구간을 형성하고, 이 신뢰구간과 영모델을 위한 판별함수와 비교해 보았다. 이는 일종의 사후검사로서 통계적 유의성과 더불어서 DIF의 정도 및 방향을 평가하는 절차이다. 만약 이 신뢰구간이 대응변수의 대부분 범위에서 영모델을 포함한다면, 이 문항은 실제적으로 심각한 수준은 아니므로 DIF문항으로 분류하지 않았다 (Miller & Spray, 1993).

질적분석

16PF 검사 문항 중 미국과 한국 집단에 대해서 차별적으로 기능하고 있는 문항들의 원인을 탐색하기 위하여 질적인 방법을 실시하였다. 우선 각 문항에 대한 DIF 분석 결과를 알려주지 않은채 두 명의 평가자들에게 영어와 한글로 된 문항을 보여주고, 문항을 평가하도록 하였다. 이 평가자들은 DIF 분석과 관련된 측정학적 지식이 있는 언어검사(language testing) 관련 전공자들로서 미국에 10년 이상 거주한 학생들로 비교적 미국과 한국 문화에 모두 익숙한 편이었다. 각 평가자는 문항을 평가할 때 구조화된 문항평가질문지(그림1 참고)를 사용하였다. 이 질문지는 Allalouf et. al (1999)가 사용했던 포맷을 부분 수정하여 마련된 것으로, 번역의 질, 차별기능문항의 방향, 그리고 원인에 대한 질문을 포함하고 있다. 구체적으로 먼저 각 평가자는 독립적으로 모든 문항들에 대해 문항평가질문지를 작성하도록 하였다. 그런 후 두 명의 평가자들과 본 연구자는 함께 모여서 그들의 평정을 설명하고 토의하도록 하였다. 이런 협의 과정 중에 평가자들은 DIF의 통계적 분석결과를 참고하였고, 필요하다면 자신의 평정을 수정

평가자 이름: _____ **문항 번호:** _____

1. 영어와 한글로 된 문항은 서로 의미상 비교 가능하다고 생각합니까?
   1) 의미상 어떠한 변화도 없다
   2) 의미상 약간의 변화가 있다
   3) 의미상 주요한 변화가 있다
2. 이 문항의 번역 수준은 어떻다고 생각합니까?
   1) 매우 우수 2) 비교적 괜찮음 3) 매우 불량
3. 이 문항에서 차별기능이 나타날 것이라고 생각합니까?
   1) 예 2) 아니오

***** 만약 당신이 1번과 3번에서 차별기능문항을 예상했다면, 다음 질문들에 응답하시오.***

4. 차별기능문항의 양이 어느 정도 일 것이라고 생각합니까?
   1) 매우 큼 2) 중간정도 3)매우 작음
5. 미국집단과 한국집단 중 어느 집단이 이 문항에 대하여 더 긍정적으로 반응할 것이라고 생각합니까?
6. 이 문항에서 나타나는 차별기능의 원인이 무엇이라고 생각합니까?
   1) 번역의 어려움
   2) 단어들이 서로 다른 난이도를 가지고 있음
   3) 문화적 적절성 문제 4) 기타 ( )
7. 이 문항을 검토한 후 추가적인 당신의 의견을 쓰시오.

그림 1. 문항평가질문지(Item Review Questionnaire). Allalouf, Hambleton, & Sireci, 1999 참고.

할 수 있도록 하였다. 최종적으로 문항평가질문지 내용을 분석하여서 차별기능문항의 가능한 원인을 밝혀냈다.

연구결과

기술 통계치(Descriptive Statistics)

척도별 신뢰도 계수를 각 인종집단별로, 그리고 각 집단내의 성별로 살펴보았다. 미국표본에서는 .78의 중앙치를 가지고, .69-.87의 내적 일관성 계수 범위를 보였고, 한국표본에서는 .62의 중앙치와 함께 .54-.87의 범위를 가지고 있었다. 미국표본의 신뢰도 계수는 16PF 검사 매뉴얼에 제시된 신뢰도 계수값 (범위 .66-.86)들과 상당한 일치를 보여 주고 있지만, 한국표본에서는 비교적 낮은 수준의 내적 일관성을 보이고 있었다. 특히, 한국 남성표본의 신뢰도 계수가 낮은 것으로 나타났다.

각 인종 집단별로 각 척도의 평균과 표준편차를 살펴보면, A척도의 경우, 미국 집단(M=16.31, SD=4.26)의 경우가 한국 집단(M=13.77, SD=3.81)의 경우 보다 더 높은 평균과 표준편차를 나타냈다. F척도의 경우 역시 미국 집단 (M=14.27, SD=4.30)이 한국 집단(M=11.76, SD=4.00) 보다 높은 평균을 보였다. 반면 H, N, Q2 척도에서는 미국과 한국 집단에서 거의 유사한 수준의 평균과 표준편차를 보였다[3].

일차원성 (unidimensionality)

5개의 외향성 척도에서 첫 번째 요인의 설명된 분산의 양을 살펴보면, 두 집단 모두에서 나머지 요인들에 의해서 설명된 분산의 양보다 충분히 큰 것으로 나타났다. 5개의 하위척도에서 미국집단의 경우 첫 번째 요인은 전체분산의 34.6%-60.8%을 설명하고 있었고, 한국표본의 경우 23.4%-59.6% 정도의 분산을 설명하고 있었다. A, F, N 척도에서 집단간 설명된 분산의 양을 비교해 보면 한국표본에서 비교적 작은 설명량을 보이지만, 대체적으로 첫 번째 요인에 의해서 설명된 분산이 5개의 척도에서 모두 충분히 크다고 볼 수 있었다. DIMTEST의 결과는 한국과 미국 집단 모두에서 일관적으로 나타났다. A, F, H, N, Q2등 5개의 하위척도들은 각각 차원상 서로 구별되는 것으로($p < .05$), 즉, 각 하위척도는 본질적인 일차원성(essential unidimensionality) 가정을 만족하였다.

인종 집단에 따른 차별기능문항 탐색

각 하위척도의 점수를 대응변수로써 사용하였고, DIF 여부를 위해 평가되는 문항의 점수는 대응변수의 점수에서 제외시켰다. 유의도 수준은 .01을 사용하였고, LDFA에서 얻어진 결과는 <표 1>부터 <표 5>에 요약되어져 있다. 먼저, A척도(표 1 참고)에서는 3개의 문항을 제외한 나머지 모든 문항이 비일방적 DIF를 나타내고 있었다. <부록>에 제시된 그림은 문항 33번에 대한 LDFA 그래프로서 비일방적인 DIF를 예시하고 있다. <부록>의 그림에서 문항 점수가 2점인 3번째 그래프를 살펴보면 완전모형의 신뢰구간이 거의 영모형을 포함하고 있는 경우로서, 이 점수에서는

3) H척도: 미국(M=10.76, SD=6.40), 한국(M=10.45, SD=6.13), N척도: 미국(M=10.56, SD=5.26), 한국(M=11.37, SD=4.14), Q2척도: 미국(M=7.59, SD=5.11), 한국(M=7.62, SD=4.55)

의미있는 DIF가 나타나고 있지 않았다. 반면 문항점수가 0점과 1점인 그래프를 보면, 영모형이 유의하게 완전모형의 신뢰구간으로부터 이탈되어 있어서 의미있는 수준의 DIF를 나타내고 있다. 그래프 상에서 DIF의 방향을 살펴보면, 문항점수가 0, 1점인 경우 모두 중간 이하의 점수 범위에서는 완전모형이 영모형보다 위에 위치함으로써 미국집단이 0과 1로 응답할 확률이 높아진다. 반면 중간점수 이상의 범위에서는 영모형이 완전모형보다 위에 위치함으로써, 한국집단에 유리하게 기능하는 비일방적 DIF를 보여주고 있다. 따라서 낮은 정도의 온정성을 갖고 있는 경우는 미국인들이 한국인들에 비하여 덜 온정적인 방향으로 응답하였고, 높은 온정성을 갖고 있는 경우는 한국인들이 덜 온정적인 방향으로 응답하였음을 알 수 있다.

표 1. 온정성 요인 (Factor A) 문항 분석 결과

문항 번호	통계적 결과 (DIF)	질적인 분석 결과				
		A	B	총평	DIF 방향	DIF 원인
1	비일방	DIF	DIF	DIF	미국	문화적
31						적절성
33	비일방	DIF	DIF	DIF	미국	번역
96	비일방					
127	비일방					
159	비일방					
63		DIF	DIF	DIF	미국	번역
65	비일방	DIF		DIF	한국	번역
98	비일방	DIF	DIF	DIF	한국	번역
129	비일방					
161						

주. 총평은 두 평가자(A, B)의 협의에 의해서 최종적으로 얻어진 평가

F척도(표 2 참고)에서는 2개의 비일방적 DIF 문항과 1개의 일방적 DIF 문항으로 10개 문항 중 3개의 문항이 DIF로 추출되어졌다.

표 2. 쾌활성 요인(Factor F) 문항 분석 결과

문항 번호	통계적 결과 (DIF)	질적인 분석 결과				
		A	B	총평	DIF 방향	DIF 원인
6						
39						
68	(한국)					
100	비일방	DIF	DIF	DIF	미국	번역
103			DIF			
134						
164						
4						
37	비일방					
70						

주. 일방 (유리하게 기능하는 집단); 총평은 두 평가자의 협의에 의해서 최종적으로 얻어진 평가임

여기서 일방적 DIF를 보인 68번 문항은 한국집단에 대해서 유리하게 기능하고 있었는데, 즉 동일한 정도의 쾌활성을 가진 미국인과 한국인이 있다고 할 때, 이 문항에 대해서는 한국인이 더욱 쾌활한 방식으로 응답하는 편파를 보일 것이다. H척도(표 3 참고)를 살펴보면 오직 한 문항 (73번 문항)이 일방적 DIF 문항으로 추출되었다. 이 문항 역시 한국집단에 대해서 유리하게 기능 하는데, 동일한 정도의 사회적 대담성을 가지고 있다고 할지라도 한국인이 이 문항에 대해서는 더욱 사회적으로 대담한 방식으로 응답할 것이다. N척도(표 4 참고)에서는 143번 문항을 제외한 나머지 9문항 모두가 비일방적 DIF 문항으로 LDFA에 의해서 추출되어졌다. 반면 Q2 척도

표 3. 사회적 대담성 요인 (Factor H) 문항 분석 결과

문항 번호	통계적 결과 (DIF)	질적인 분석 결과 A	B	총평	DIF 방향	DIF 원인
9						
73	(한국)		DIF	DIF	한국	번역
135						
137						
41						
71		DIF		DIF	미국	번역
105			DIF	DIF	한국	번역
107						
167						
169						

주. 일방 (유리하게 기능하는 집단); 총평은 두 평가자의 협의에 의해서 최종적으로 얻어진 평가임

표 4. 내밀성 요인 (Factor N) 문항 분석 결과

문항 번호	통계적 결과 (DIF)	질적인 분석 결과 A	B	총평	DIF 방향	DIF 원인
47	비일방					
50	비일방					
80	비일방					
113	비일방					
143						
148	비일방					
15	비일방		DIF	DIF	한국	번역
18	비일방					
84	비일방	DIF	DIF	DIF	미국	번역
117	비일방					

주. 총평은 두 평가자의 협의에 의해서 최종적으로 얻어진 평가

(표 5 참고)에서는 121번, 123번과 156번 등 세 개 문항이 일방적 DIF 문항으로 밝혀졌다. 이 중 121번 문항은 한국 집단에 대해서, 나머지 2문항은 미국 집단에 대해서 유리하게 기능하고 있었다. 예컨데, 동일한 정도의 자기의존성을 갖고 있다고 할 지라도 이 156번 문항에 대해서는 미국인이 한국인에 비하여 더 자기 의존적인 경향으로 응답하는 편파를 보일 것이다. 요약해 보면, 51개의 외향성 문항 가운데 24개 문항이 두 집단에 대해서 차별적으로 기능하고 있는 것으로 나타났다. 각 척도별로 볼 때, A와 N 척도는 70% 이상의 문항들이 미국과 한국집단에 의해서 동등한 의미로 해석되고 있지 않음이 밝혀졌다.

표 5. 자기의존성 요인 (Factor Q2) 문항 분석 결과

문항 번호	통계적 결과 (DIF)	질적인 분석 결과 A	B	총평	DIF 방향	DIF 원인
27						
59						
89						
121	(한국)	DIF	DIF	DIF	미국	번역
152						
25						
56						
92						
123	(미국)					
156	(미국)		DIF	DIF	미국	번역

주. 일방(유리하게 기능하는 집단); 총평은 두 평가자의 협의에 의해서 최종적으로 얻어진 평가임

질적 분석

질적인 분석 결과는 <표 1>부터 <표 5>에 통계적 분석 결과와 함께 제시되어 있다. 두 명의 평가자들로부터 공통적으로 차별기능을

보인다고 평정된 문항의 수는 총 13개로써, (a) A 척도에서 5개 문항, (b) F 척도에서 1개 문항, (c) H 척도에서 3개 문항, (d) N 척도에서 2개 문항, 그리고 (e) Q2 척도에서 2개 문항이었다.

질적 분석 Vs. 통계적 분석

51개 문항 중 24개의 DIF 문항을 밝혀낸 LDFA의 결과와 비교해 볼 때, 질적인 분석이 훨씬 적은 수의 문항에서 DIF를 발견하였다. 질적분석에서 DIF로 평정된 문항들 가운데 세 문항(63번, 71번, 105번)을 제외한 나머지 10 문항은 LDFA에 의해서 DIF로서 평가되었다. 그러나 <표 1>과 <표 3>에서 보듯이 63, 71, 105번 문항들은 통계적으로는 유의미하게 차별기능을 보이지 않았으나, 평가자들에 의해서 두 언어집단에서 동등하지 않은 문항이라고 평정되었다. 반면 LDFA에 의해 DIF를 나타낸다고 밝혀진 24개 문항 중 14개의 문항들에 대해서 두 명의 평가자들은 어떠한 차별기능도 나타내지 않고 두 집단 모두에 동등하게 기능하고 있다고 평정하였다.

차별기능문항의 원인

평가자들은 차별기능문항의 원인을 크게 두 가지로 즉, (a) 번역의 어려움 혹은 번역 실수, (b) 문화적 적절성 문제로 분류하였다. 평가자들에 의해서 차별기능을 보인다고 평정된 13개 문항 중 오직 한 개 문항만이 문화적 적절성 문제로서 분류되었고, 나머지 12개 문항은 번역의 문제로서 평가되었다.

번역의 문제: 이 연구에서 발견된 번역의 문제를 보다 세부적인 유형으로 (a) 단어 뉘앙스의 적절성 문제, (b) 숙어적 표현, 그리고 (c) 문장 양태 (Mode of Expression)로 분류할 수 있었다. 먼저 가장 어려운 번역의 문제로서 일부 특정한 단어들의 뉘앙스를 잘 살려서 적절하게 번역하는 문제이다. 예를 들어서 평가자들은 33번 문항의 people's needs (사람들의 요구), 63번의 affection 혹은 caring (애정이나 남을 돌보는), 그리고 65번의 sales manager (판매 관리자)와 같은 단어들의 번역 문제를 지적하였다. 또한 쾌활성을 측정하는 100번의 문항은 "I like to go out to shows or entertainment often."으로 이를 한글판에서는 "나는 쇼나 오락을 종종 보러간다"라고 단어별 일대일 해석을 하였다. 하지만 "shows"나 "entertainment"는 "쇼"나 "오락"이라는 직역보다는 구체적인 단어로서 번역되는 것이 적절하다고 평가되었다. 대체로 이렇게 직역된 단어들은 한국사람들에게 분명한 의미를 전달하기 힘들기 때문에, 라이브 콘서트, 영화관, 혹은 노래방 등 보다 구체적으로 번역되는 것이 바람직 할 것이라고 평가되었다. 두 번째 번역의 문제는 숙어적 표현이다. 예를 들어서 "put all your cards on the table" 혹은 "play your hand close to your chest" (84번 문항)과 같은 숙어는 우리나라 말에 대응되는 적절한 숙어적 표현이 없기 때문에 비숙어적인 표현으로 완곡하게 번역되었다. 이처럼 숙어적 표현을 비숙어적 표현으로 번역하는 과정에서 정확하게 그 원래의 의미를 전달할 수 없었을 지도 모른다. 마지막으로 71번 문항 등에서 발견된 문장 양태의 변화에 기인한 번역 문제이다. "I tend to get embarrassed if I suddenly become the center of attention in a social group"이라는 문장이 한글판에서는

"만약 내가 갑자기 사교 모임에서 주목을 받게 된다면 당황스러울 것이다"라고 번역되었다. 영문판에서는 특정한 상황에서 사람들의 경향성을 묻는 것이라면, 한글판에서는 가정법의 양태로 번역되었다. 평가자들의 토의결과에 따르면 미국인들은 이 문항에 대하여 자신의 경험에 근거하여 대답을 할 것이고, 한국 사람들은 상상이나 추측에 의하여 대답했을 가능성이 높기 때문에 이 문항이 두 집단에서 동등하게 해석되지 못했을 가능성을 지적하였다.

문화적 적절성 문제: 또 다른 차별기능문항의 원인은 문화적 적절성 문제였다. 이 경우는 번역상 어떠한 문제도 발견되지 않았으나 일부 문항의 문화적 내용 때문에 두 집단이 서로 다르게 반응한다는 점이다. 즉 번역된 문장이 영어판 문장과 비교하여서 동일한 함축적 (implicative) 의미를 전달하지 못하는 경우이다. 예를 들어서, 온정성을 측정하는 1번 문항의 경우 "I'd enjoy more being a counselor than being an architect" 라는 문장은 "나는 건축가가 되느니 상담자가 되는 게 더 좋겠다"라고 번역되었고, 번역상으로는 어떠한 문제도 발견되지 않았다. 그러나 상담자라는 직업은 미국 사회와 비교해 볼 때, 한국인들에게는 비교적 익숙하지 않은 직업이고, 반면 건축가라는 직업은 한국 사회에서 상당히 선호되는 직업이다. 따라서 사람들을 만나고 인간지향적인 직업인 상담자와 혼자서 주로 작업하는 직업인 건축가를 비교하여 응답자의 외향성을 측정하려던 본래의 문항 목적과는 상관없이 이 번역된 문항은 다른 내용을 측정할 수도 있다는 점이다. 예를 들어, 한국인들은 그들의 외향성 정도와는 상관없이 직업의 수입이나 장래성을 고려하여 상담자 보다는 건축가를 더 선호하는 방식으로 응답했을 가능성도 배제할 수 없다.

논 의

현재 우리 나라의 경우 외국에서 개발되어 그 타당성을 인정받은 많은 심리 및 교육검사들을 한글로 번안하여 사용하고 있다. 이 경우 반드시 연구자들이 고려해야 할 점은 검사가 측정하고 있는 구인에 대한 타당도 검증 문제이다. 그 구인이 우리 나라 언어 및 문화에서 동일한 방식으로 이해되고 있는 것인지에 대한 검토는 항상 필요하다. 대략적인 검사 번안 절차는 다음과 같이 요약해 볼 수 있다. 일단 검사 번안의 필요성에 대하여 신중히 검토하고, 그 후 검사 번역자와 번역절차에 대한 결정을 해야 한다. 그런 후 검사 번안을 시행하고, 적절한 문항 수정이 이루어질 것이다. 더불어서 앞으로 그 검사도구가 어떤 분야에 어떤 방식으로 사용될 것인지에 대해서 생각해 보아야 한다. 또한 문항편파연구, 요인분석법, 준거관련 연구와 같은 타당도 증거를 수집할 수 있는 연구를 계획하여야 한다. 마지막으로 모든 번안과정과 타당도 증거에 대한 문서화 역시 중요한 과정이다.

본 연구에서는 성격검사 문항의 동등성을 확립 및 평가하기 위하여 판단적인 방법과 통계적인 방법을 상호보완적으로 사용하였다. 먼저 선번안 기법을 이용하여 한글판 16PF 검사 문항을 마련하였다. 그런 후 문항 수준에서의 동등성을 평가하기 위하여 로지스틱 판별분석법을 이용한 DIF 분석을 실시하였다. 더불어서 각 문항에 대한 질적인 분석을 보

완적으로 실시하여, DIF의 원인에 대한 탐색을 시도하였다. 통계적 분석 결과와 비교해 볼 때, 평가자들에 의한 질적 분석은 상대적으로 적은 수의 차별기능문항을 보고하였고, 통계적 분석 결과 차별적으로 기능한다고 판단된 문항 중 14개 문항은 평가자들에 의해서는 두 집단에서 동등하게 기능하고 있는 문항이라고 판단되었다. 두 분석으로부터 나온 이런 불일치된 결과는 몇 가지의 이유로 설명해 볼 수 있다. 먼저, 통계적 방법에 의해서는 두 가지 종류의 차별기능문항, 즉 일방적 혹은 비일방적 DIF 모두를 판별해 내지만, 질적 분석인 경우는 오로지 일방적인 DIF만을 논의하게 된다는 점에서 기인할 수 도 있다. 또 다른 설명으로는 평가자들이 차별기능문항의 가능한 원인이라고 파악한 번역이나 문화적 적절성 문제 이외에 질적인 분석에서 파악해 낼 수 없었던 또 다른 가능한 DIF 원인들이 존재할 수 있다는 점이다. Scheuneman (1984;1987)이 논의하였듯이, 차별기능문항이란 문항과 응답자들간의 복잡한 상호작용의 결과이기 때문에, 어떤 단일한 차별기능문항의 원인은 존재할 수 없다. Shealy와 Stout (1993)의 차별기능문항에 대한 이론적 구조에 의하면 하나의 검사는 하나의 목표능력(혹은 일차적 능력)으로 구성되어 있지만, 문항 반응들이 하나 혹은 그 이상의 잡음 결정인자(혹은 이차적 능력)에 의하여 결정될 때 차별기능문항에 대한 잠재성이 나타난다고 정의하였다. 이러한 잡음 결정인자 혹은 이차적 성향들이 질적인 분석을 통하여 밝혀질 수도 있지만, 여러 가지 요인들이 상호작용 하여 나타나는 경우에는 그것들에 대한 파악이 힘들어 질 수도 있다. Gierl 외 연구자들은(1999) 피험자들이 각 문항에 응답하면서 사용하는 인지적 과정이 여러 언어판 검사문항을 응답할 때 서로 다르게 작용할 수 있다고 지적하며, 이러한 심리적인 요인들을 가능한 차별기능문항의 원인으로 제안하였다. 따라서 집단간 인지와 검사수행간의 관계에 대한 심리적인 요인들을 보다 체계적으로 연구함으로써 차별기능문항에 대한 이해를 증진시킬 수 있을 것이라고 제안하였다.

번안된 성격요인검사 문항에서 나타난 차별기능문항의 원인을 질적으로 탐색해 본 결과 그 가장 큰 원인은 번역상의 어려움으로 나타났다. 비록 두 명의 평가자들이 51개 문항들의 번역의 질에 대하여 일반적으로 우수하다고 평가하였지만 일부 문항에 대해서 번역상의 어려움이 나타났고, 이러한 문항들이 두 집단에서 차별적으로 기능한다고 판단되어졌다. 이는 번역된 학업 성취도와 학업 적성검사를 다루었던 선행연구의 결과와 일치하는 것으로 외국의 검사를 번역하여 사용하는 경우 번역절차의 중요성을 시사하고있다 (Allalouf et. al., 1999; Gierl et. al., 1999). 일반적으로 검사를 번역해서 사용하는 경우, 인지적 검사들에 비해서 성격검사와 같은 비인지적 검사에서 문화적 편파 경향이 나타날 가능성이 높다고 알려져 있다(Reynolds & Brown, 1984). 이는 문항내용의 문화적 적절성 문제로서 일부 내용이 특정한 집단에게는 생소한 것으로 번역된 문항이 原문항에서와 같은 함축적 의미를 전달할 수 없다는 문제이다.

본 연구로부터 얻어진 경험을 바탕으로 앞으로 외국의 검사를 차용하려고 하는 연구자들이나 번역된 문항에서 차별기능문항의 가

능한 원인을 평가하는 연구를 위한 몇 가지 방향을 제시하고자 한다. 먼저 본 연구의 한 계점으로 오직 두 명의 평가자들을 이용하여 질적인 분석을 시도했다는 점을 들 수 있다. 두 명의 평가자들이 언어검사를 전공하고 있는 학생들로서 언어와 문화 뿐 아니라 측정학적 개념에 대한 지식이 있는 전문가였다고 할 지라도 보다 체계적이고 폭넓은 질적 분석을 위해서는 두 명 이상의 전문가들이 필요하다는 점이다. 적어도 6명-10명 정도로 구성된 다양한 배경을 가진 전문가 집단을 형성하여 몇 차례에 걸친 토론이 필요하며, 효과적인 문항 평가를 할 수 있는 체계적인 훈련과정도 요구된다.

이 연구의 또 다른 제한점으로는 문항 동등성을 평가하는 통계적 기법으로 단 한 개의 차별기능문항 분석법만이 사용되었다는 점이다. Pontenza와 Dorans(1995)에 의하면 적어도 네 가지 차원으로 차별기능문항 분석법들을 분류할 수 있고, 이처럼 서로 다른 기법들은 문항 수행을 수량화하고 대응변수를 결정하는데 있어서 서로 다른 방법들을 사용한다. 따라서 이들은 DIF에 관한 서로 다른 정보를 제공하는 경향이 있으므로 가능하면 차별기능문항 분석법을 선택할 때 서로 다른 접근법을 가진 두 개 이상의 방법을 상호보완적으로 사용하는 것이 바람직할 것이다. LDFA는 표본의 크기가 적절히 큰 경우 상당히 좋은 수행을 보이지만, 두 집단의 능력분포가 동일하지 않은 경우에는 일종오류의 가능성이 커질 수 있다는 단점을 지니고 있다 (Miller & Spray, 1993). 이런 경우, 능력 분포가 동일하지 않아도 상당한 항내성을 가지고 있는 차별기능문항 분석법, 예를 들어 Shealy와 Stout(1993)의 SIBTEST를 보완적으로 사용한다면 번안된 검사에서 나타나는 DIF에 대한 이해를 보다 증진시킬 수 있을 것이다.

세 번째로는 우리 나라 현실에서 번안된 검사 문항을 효율적으로 평가하고, 차별기능문항의 원인을 파악할 수 있는 일종의 지침의 개발이 필요하다는 점이다. Allalouf 외 연구자들은(1999) 인지적 검사인 언어 검사를 대상으로 차별기능문항의 원인을 연구하고 그 결과를 바탕으로 하나의 플로어 차트를 개발하였다. 이는 일종의 차별기능문항의 원인을 파악하는 과정을 설명하는 것으로 가장 기본적인 번역의 문제에서부터 보다 복합적인 수준인 문화적 적절성 문제로 이르는 차트를 제안하였다. 이 차트에서 우선 (1) 번역이 올바른지?, (2) 문장형태가 동일한지?, (3) 단어의 난이도가 동일한지?, 그리고 (4) 문화적 적절성에서 어떠한 차이가 있는지? 순으로 질문을 하고, 이에 따라서 차별기능문항의 원인을 파악해 나간다. 이는 히브리어와 러시아어 문항들을 비교하여 얻은 결과로서 이것이 실제로 다른 나라 언어 특히 한국어에도 일반화시킬 수 있는 과정인지 또한 성격검사와 같은 비인지적 검사를 검토할 때도 적용가능한 것인지에 대해서는 충분한 연구가 필요할 것이다. 현재 우리 나라 교육학이나 심리학의 많은 연구들이 외국의 검사를 번역해 사용하고 있는 실정을 감안한다면 앞으로 한국의 실정에 맞는 적절하고 체계적인 지침 개발은 필수적일 것이다.

다음으로 외국어 검사를 한글로 번역할 때 사용할 수 있는 효과적인 번역기법 개발이 필요하다는 점이다. 본 연구나 기존 선행 연구들의 결과에서도 밝혀졌듯이 대다수의 차

별기능문항의 원인은 번역의 문제 혹은 어려움으로 나타났다. 따라서 검사 번역의 초기과정에서 보다 체계적이고 효과적인 방법을 이용하여 번역을 시도한다면 각 언어와 문화간 동등한 검사개발의 기초가 될 것이다. 예를 들어서 본 연구에서는 선번안 기법을 응용하였는데, 후속연구로서 동일한 자료에 대해 역번안 기법을 적용해 보고, DIF 분석을 실시하여 그 결과를 비교해 본다면, 두 번역기법의 효율성에 대한 경험적인 검토가 가능할 것이다. 이처럼 앞으로 여러 가지 다양한 번역 기법들의 효과성을 검토하는 경험적 연구들이 진행되어야 하며 이를 통해 외국의 검사를 번역하여 자신들의 연구에 사용하려는 연구자들에게 도움을 줄 수 있을 것이다.

마지막으로, 검사들간의 동등성 확립은 단일한 하나의 방법에 의해서 이루어 질 수 없는 것이고, 다양한 방법을 필요로 하는 연속적인 과정이라는 점을 강조하고 싶다. 우선, 검사를 번안하는 과정에서 상당한 시간과 노력이 투자되어야 하며, 일단 번안된 검사를 비교문화연구에 사용하기에 앞서서 통계적인 방법과 질적인 방법을 이용하여 다각적으로 검사의 동등성을 검토해보아야 할 것이다.

참 고 문 헌

김신영(1993). 문항편파성의 원인탐색 연구. 한국교육, 20, 199-213.

김아영. 임은영 (2003). Effects of different types of practice in cross-cultural test adaptation of affective measures. *Korean Journal of Psychology: General. 22(1)*, 89-113.

염태호, 김정규 (1990). *성격요인검사: 실시요강과 해석방법*. 서울: 한국심리적성연구소.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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.

Brislin, R. W. (1970). Back translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*, 185-216.

Butcher, J. N. (1996). *International Adaptations of the MMPI-2: Research and Clinical Applications*. Minneapolis, MN: University of Minnesota Press.

Casagrande, J. B. (1954). The ends of translation. *International Journal of American Linguistics, 20*, 335-340.

Cattell, R. B., & Cattell, H. E. (1995). Personality structure and the new fifth edition of the 16 PF. *Educational and Psychological Measurement, 55(6)*, 926-937.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95(1)*, 134-145.

Drasgow, F., & Probst, T. M. (2000). Evaluating measurement equivalence across languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Erlbaum.

Ellis, B. B., & Mead, A. D. (1998, August). *An application of the DFIT framework to assess the measurement equivalence of a Spanish translation of the 16PF questionnaire*. Paper

presented at the annual meeting of the International Congress of Applied Psychology, San Francisco, CA.

Gierl, M. J., & Khaliq, S. N. (in press). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement.*

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Using statistical and judgmental reviews to identify and interpret translation DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Candada.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9(1),* 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A Progress Report. *European Journal of Psychological Assessment, 10(3),* 229-244.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11(3),* 147-157.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of scale translations. *Journal of Applied Psychology, 67,* 818-825.

Hulin, C. L., Drasgow, F., & Parson, C. K. (1983). *Item response theory: Application to psychological measurement.* Honewood, IL: Dow Jones Irwin.

Jöreskog, K., & Sörbom, D. (1986). *PRELIS 2: Users reference guide.* Chicago: Scientific software international.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107-122.

Poortinga, Y. H., & Van de Vijver, F. J. R. (1987). Explaining cross-cultural differences: Bias analysis and beyond. *Journal of Cross-Cultural Psychology, 18,* 259-282.

Potenza, M. T., & Dorans, N. J. (1995) DIF assessment for polytomously scored items: a framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23-37.

Reynolds, C. R., & Brown, R. T. (1984). Bias in mental testing: an introduction to the issues. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing.* New York: Plenum Press.

Russell, M. & Karol, D. (1994) *16PF Fifth Edition: Administrators manual.* Champaign, IL: The Institute for Personality and Ability Testing, Inc.

Scheuneman, J. D. (1984). A theoretical framework for the exploration of causes and effects of bias in testing. *Educational Psychologist. 19(4),* 219-225.

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement, 24(2),* 97-118.

Shaughenssy, M. F., & Kang, M. H. (1998). *Personality profile of gifted children: The 16PF Fifth Edition A Comparative study of Korean and US Children.* Unpublished manuscript.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Sireci, S. G., Bastari, B., & Allalouf, A. (1998, August). *Evaluating construct equivalence across adapted tests.* Paper presented at the annual meeting of the American Psychological Association,

San Franscisco, CA.

Sohn, W. (2002). Equivalence of Constructs Measured by Two Different Language Versions of 16PF. *Korean Journal of Psychology: General, 21(1)*, 91-116.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589-617.

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

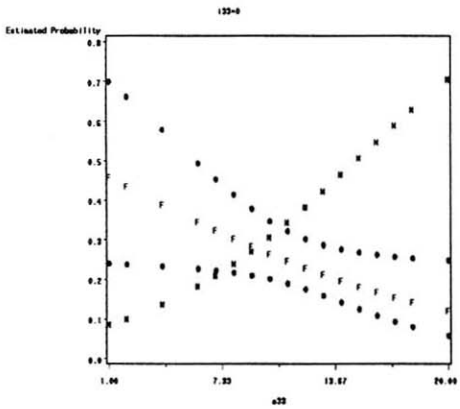
van de Vijver, F., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47(4)*, 263-279.

Werner, O., & Campbell, D. (1970). Translating, working through interpreters, and the problem of decentering. In R. Naroll and R. Cohen (Eds.), *A handbook of methods in cultural anthropology*. New York: American Museum of Natural History.

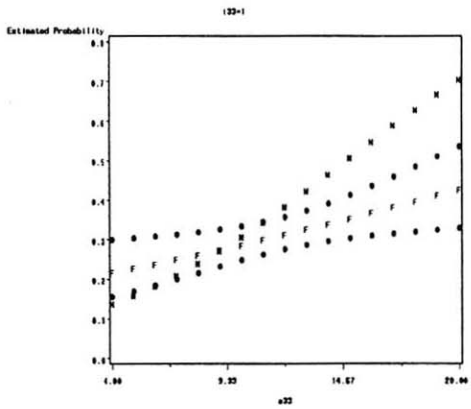
1 차원고접수일 : 2003. 3. 24
수정원고접수일 : 2003. 11. 24
최종게재결정일 : 2003. 12. 2

부 록

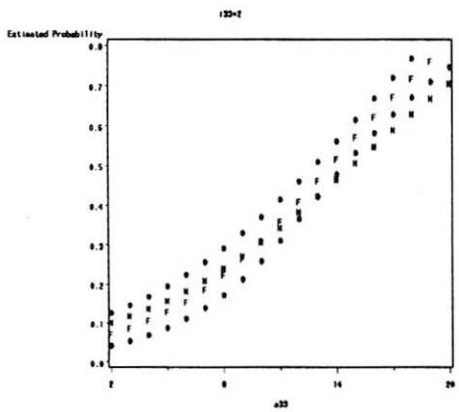
1. 문항점수=0



2. 문항점수=1



3. 문항점수=2



* 그림 설명

[F]: 완전모형 =P(G=미국|A요인, 문항33)

[N]: 영모형 =P(G=미국| A요인)

[●]: 95% 신뢰구간

한국심리학회지: 일반
The Korean Journal of Psychology
2003. Vol. 22, No. 2, 57-80

A Comprehensive approach for adapting psychological tests

Wonsook Sohn
Korea Institute of Curriculum & Evaluation

The main purpose of this study was to describe a comprehensive approach for empirically adapting psychological tests for use in multiple languages and cultures. In particular, this study focused on how the statistical and judgmental methods complement each other to establish cross-cultural equivalence. Another focus of the study was how differential item functioning (DIF) research can be best extended to the problem of evaluating the equivalence of tests across languages. Also this study introduced guidelines for adapting educational and psychological tests developed by the International Test Commission(ITC), which are widely used in many other countries. Finally it was emphasized that much attention should be given to the preparation of guidelines on adapting tests and test use appropriate for Korean situation.

Key Words: Test adaption, Score Equivalence, Differential Item Functioning