차별문항기능 기법의 응용 : 교육 및 심리검사의 번안과정에서

손 원 숙*

이 연구의 주된 목적은 외국의 교육 및 심리검사를 번안하여 사용하는 경우, 번안된 검사와 原 검사간의 평형성 (equivalence)을 경험적으로 측정하기 위하여 차별문항기능기법을 응용하는 것이다. 즉, 평형성이라는 것은 타당한 그룹비교를 위한 하나의 전제조건으로서 이런 평형성이 확립되지 않은 채 번안된 검사들의 무분별한 사용은 바람직하지 않다. 서로 다른 접근법을 가지고 있는 두 개의 차별문항기능기법, 로지스틱 판별분석과 SIBTEST가 상호보완적인 목적으로 이용되었고, 최선의 대응변수를 모색하려는 시도로서 다변량 대응기법과 순환적인 정화기법이 사용하였다. 5개의 하위척도를 가진 외향성 척도 중 3개의 척도에서는 적은 수의 차별기능문항을 발견하였으나, 나머지 2개의 척도에서는 반 이상의 문항에서 차별기능을 추출하였다 마지막으로, 대응변수의 선택에 대한 문제들이 논의되고 있다.

I. 서 론

교육 및 심리검사들을 살펴보면, 하나의 검사를 다양한 언어와 문화에 맞게 번안(*translating or adapting*)하여 사용하고 있는 예를 흔히 볼 수 있다. 대체적으로 연구자들이 한 검사를 여러 가지 언어로 번안하여 사용하는데 관심이 있는 이유는 몇 가지로 정리해 볼 수 있다. 첫째, 비교문화 연구자들의 기본 관심사중 하나인 심리학적 개념의 보편성 혹은 특수성을 평가해 보고자 하는 것이다. 즉, 서로 다른 문화 속에 살고 있는 사람들의 유사성 혹은 차이점에 관심이 있는 경우이다. 두 번째로는 국제 학력 비교연구에 대한 관심의 급증으로, 국제교육성취평가협회(International Association for the Evaluation of Educational Achievement)가 실시한 제 3차 (1995년)와 4차(1999년)

* 이화여대 심리학과

국제 수학 및 과학 연구 (TIMSS)가 최근의 한 예라고 볼 수 있다. 또한 검사 번안은 경제적인 이유와 간편성이라는 장점 때문에 쉽게 이루어 질 수 있다. 한 연구자가 어떤 하나의 특별한 구성개념(construct)을 측정하고자 하지만 자국의 언어로 된 검사가 없고, 새로운 검사를 제작하여 사용할 만한 시간과 비용이 없는 경우 흔히 다른 나라에서 개발된 검사를 차용해서 사용하는 경우가 많다. 특히 이러한 연구들은 다민족으로 구성된 미국, 캐나다, 그리고 유럽에 있는 국가들에서 관심이 고조되고 있는 추세이며, International Test Commissions (ITC)에서는 최근 교육 및 심리검사를 번역 혹은 번안하는데 참고할 수 있는 지침서를 발표하였다(Hambleton, 1994). Hambleton (1993;1994)은 최근 국제학력비교나 비교문화연구에 대한 관심이 증가하고, 다양한 언어로 이루어진 자격시험에 대한 수요가 급증하고 있는 현 추세로 본다면, 이러한 경향은 앞으로도 계속될 것이라고 예상하였다.

이런 현재의 추세에도 불구하고, 연구자들이 검사를 번안하여 사용하는데 몇 가지 방법론적인 문제들이 내재되어 있다(Church & Lonner, 1998; Hambleton & Kanjee, 1995). 첫째, 점수의 평형성에 초점을 두면서 검사를 번안할 수 있는 방법들의 개발, 둘째, 비교 문화 혹은 비교 국가 자료를 사용하고 해석하는 방법들, 마지막으로 검사를 번안하는 지침의 개발과 사용에 관한 것이다. 이 중에서 첫 번째 문제는 바로 비교 문화연구의 가장 근원적인 이슈라고 볼 수 있는 "평형성(equivalence)"의 확립에 관한 것이다. "평형성"이라는 것은 검사 점수들이 서로 다른 그룹에서 비교될 수 있는 측정수준을 말하는 것으로, 이런 "평형성"이 확립이 되지 않는다면 그룹간의 타당한 점수비교가 어려울 것이다. 즉 그룹간 비교 연구를 하기에 앞서서 꼭 만족되어야만 하는 하나의 전제조건으로 이 평형성에 대한 경험적인 검토는 반드시 필요하다. Drasgow (1984)는 이런 평형성을 문항반응이론의 맥락에서 "관찰 점수와 그 검사가 측정하려고 하는 잠재속성간의 관계가 하위 그룹들에서 동일할 때 평형적인 측정(equivalent measurement)이 얻어지는 것"이라고 정의 내렸다.

최근까지 많은 비교문화연구자들은 이 평형성을 보장하기 위하여 역번역(back-translation)기법과 같은 질적인 방법을 주로 사용하였다. 이 역번역 기법은 여러 개의 경험적인 연구들에서 밝혀졌듯이 번역의 질을 평가하는 초기작업에는 상당히 유용한 것으로 나타났지만 (Hulin & Mayer, 1986; Hulin, Drasgow, & Komocar, 1982), 비교문화의 평형성(cross-cultural equivalence)을 수립하는데에는 충분하지 않으며, 이와

더불어서 통계적인 방법의 필요성이 제기되어 지고 있다(Ellis, Minsel, & Becker, 1989; Hulin, et al., 1982). ITC 지침에서도 "검사 개발자들은 여러 언어유형의 검사의 평형성을 형성하기 위하여 적절한 통계적인 방법을 체계적인 질적인 방법과 더불어서 사용하기를 권장한다"고 하였다 (Hambleton, 1994). 국내의 한 연구 (Kim & Lim, 1999)에서도 검사의 평형성을 확립하기 위한 3가지의 방법(즉, 단순번역+검토, 역번역+검토, 역번역+검토+경험적인 타당도검증)을 비교해 본 결과, 역번역기법과 경험적인 타당도 연구를 함께 사용한 경우에 가장 바람직한 결과를 산출하였다고 보고하였다.

최근에 검사의 비교문화적 평형성(cross-cultural equivalence)을 평가하기 위하여 사용할 수 있는 하나의 통계적인 방법으로 차별문항기능(Differential Item Functioning: DIF) 기법이 소개되고 있다. 대체적으로 인지적인 검사들, 예컨데 학업성취도 검사에 주로 DIF 기법이 적용되어 왔으나(성태제, 1994; 송미영, 2001), 최근에는 많은 연구자들이 이런 영역 이외에서 DIF 기법을 활용하는데 관심을 가져오고 있다. 예컨데, 비인지적인(non-cognitive) 검사들에서 동일한 잠재요인(latent trait)을 가지고 있으나 서로 다른 그룹에 속한 개인들이 검사의 문항을 동일하게 해석하고 있는지 알아보기 위하여 DIF 기법은 활용되고 있다 (Brown, 1996; Ellis, Becker, & Kimmel, 1994; Johanson & Johanson, 1996; Sohn, 1998). 또한 검사 번안 장면에서도 다양한 언어와 문화 그룹 간 문항의 평형성을 연구하기 위하여, DIF 기법을 활용하고 있다(Botempo, 1993; Budgell, Raju, & Quartetti, 1995; Drasgow & Probst, 2000; Roznowski & Reith, 1999). 이런 연구를 통하여 연구자들은 번역의 질, 잠재변인과 문항들간의 관계, 그리고 서로 다른 그룹간의 검사 점수의 평형성에 대한 정보를 얻을 수 있으며, 또한 문제를 나타내고 있는 문항들의 소재를 찾아낼 수 있다는 점에서 DIF 기법들은 상당히 유용하게 응용되고 있다. 이런 상황에서 서로 다른 언어 혹은 문화 그룹에 속하는 개인들이 검사가 재고 있는 속성, 즉 동일한 능력, 성격, 혹은 태도를 가지고 있다면 이 두 개인들은 같은 문항에 대하여 동일한 방식으로 응답하기를 기대한다. 그러나, 그들이 동일한 방식으로 반응하지 않는 다면 그 문항은 "DIF"를 나타낸다고 할 수 있으며, 그 원인에 대한 심층적인 분석이 필요할 것이다.

그러나, 검사 번역 장면에서 DIF 기법을 응용하는데 있어서, 몇가지 한계점들이 존재한다. 첫째, 서로 다른 문화 혹은 언어그룹에 속하는 개인들을 적절하게 대응시킬 수 있는 대응 변수(matching variable)를 선택하는 문제이다. 전통적인 DIF 연구에서

는 검사의 총점이 주로 대응변수로 사용되어 왔었으나, 번역 DIF 연구(translation DIF studies)에서는 이 총점이 적절치 않을 수 도 있다. 일반적으로 번역되어진 검사들이 원본의 검사에 비하여 낮은 신뢰도를 가지고 있으며, 따라서 DIF가 실제로 존재하지 않으나, DIF를 잘못 발견할 수 있는 확률, 즉 일종오류를 범할 가능성이 높아지기 때문이다(Zwick, 1990). 또한 검사를 번역하는 과정에서 두 개의 서로 다른 언어로 된 검사들을 서로 다른 난이도로 번역하는 체계적인 편파를 나타낼 수 있다는 점이다(Sireci, 1997). 따라서 검사의 총점 대신 새로운 다른 대응변수의 모색이 필요하다.

두 번째로는 기존의 많은 번역 DIF 연구들이 다분문항을 이분문항으로 양분화 시켜서 이분문항에 기초한 DIF 기법들을 사용해 왔다는 점이다 (Ellis & Mead, 1998; Ellis et al., 1989; Huang, Church, & Katigbak, 1997; Hulin & Mayer, 1986). 이런 연구들은 개념상으로 다분문항을 양분화 시키는 것이 타당하다고 가정하였거나, 대부분의 DIF 기법들이 이분 문항에 기초하여서 개발되어 졌기 때문에 다분 문항에 기초한 기법들이 그리 대중적이지 않으므로, 이런 실행들이 이루어져 왔다. 하지만 일부 문화 그룹 구성원들은, 특히 한국인들은 극단적인 반응(전혀 그렇지 않다 혹은 매우 그렇다)보다는 중립적인 반응(그저 그렇다, 잘 모르겠다, 혹은 그럴 때도 있다)을 주로 선택하는 경향이 있기 때문에 만약 다분 문항을 임의적으로 양분시킨다면, 이와 같은 문화적 차이에 대한 중요한 정보를 잃을 수도 있을 것이다.

지금까지 언급한 문제점들에 대한 어떤 절대적인 해결법은 있을 수 없겠지만, 본 연구에서는 몇 가지 시도를 통하여 어느 정도 이런 문제점들에 대해 접근해 보고자 한다. 첫째, 좀 더 정확한 대응(matching)을 위하여, 다변량 대응 기법 (multivariate matching technique)과 순환적인 정화기법 (iterative purification process)을 사용하였다. 두 번째로는 세 개의 반응 옵션을 가지고 있는 16PF 검사 문항을 위하여 다분 문항에 적합한 두 개의 DIF 기법을 이용하였다: (a) 로지스틱 판별 분석 (logistic discriminant function analysis: LDFA) (Miller & Spray, 1993), (b) Simultaneous Item Bias Test (SIBTEST) (Shealy & Stout, 1993). 서로 다른 DIF 기법들은 문항수행(item performance)을 수량화하고 대응변수를 결정하는데 있어서 서로 다른 방법들을 사용하기 때문에, 이 두 가지 기법을 상호보완적으로 사용함으로써, DIF에 대한 서로 다른 정보를 얻을 수 있을 것이다(Penfield & Lam, 2000).

Ⅱ. 차별 문항 기능 기법

이 연구에서는 외향성을 측정하고 있는 성격검사 문항들이 두 개의 문화와 언어가 다른 두 그룹에서 평형성(equivalence)을 유지하고 있는지 평가하기 위하여 다음과 같은 두 가지 차별문항기능기법을 사용하였다.

1. 로지스틱 판별 분석 (Logistic Discriminant Function Analysis)

Miller와 Spray(1993)는 로지스틱 회귀분석의 하나의 변형으로서 로지스틱 판별분석을 제안하였다. 이 판별함수는 다음과 같이 쓰여 질 수 있다.

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

여기에서 G는 소속 집단, X는 대응변수의 점수, U는 각 문항의 응답, 그리고 X*U는 X와 U, 두 개 변인의 積(product)이다. 이 판별함수의 계수는 α(i = 0, 1, 2, 3)로서 나타내고, 이들은 우도 함수(likelihood function)를 최대화 시키는 방식으로 추정된다. 여기서 소속 집단, G = 1 이면 준거(reference) 집단, G = 0이면 초점(focal) 집단을 가리키며, 각 문항의 반응, U는 두 개 이상의 범주를 모두 포함시킬 수 있어서 이분 문항 뿐 아니라, 다분 문항도 이 함수에서 다룰 수 있다. 판별 계수 α_3, α_2 의 유의성에 대한 우도비검정(likelihood ratio tests)은 비일방적 DIF (nonuniform)와 일방적 DIF (uniform)를 평가하는 것으로, 만약 $\alpha_2 \neq 0$ 이고, $\alpha_3 = 0$ 이면, 그 문항은 일방적 DIF를 나타내는 것이고 $\alpha_2 = 0$ 이고, $\alpha_3 \neq 0$ 이면, 그 문항은 비일방적 DIF를 나타내는 것이다. 구체적으로, 이 로지스틱 판별분석에서는 세가지의 모델이 탐색되어지는데 첫 번째, 완전모형(full model)은 대응점수, 문항, 그리고 대응점수와 문항의 상호작용으로 구성되어 있고 (공식 1), 두 번째, 축소 모형(reduced model)은 완전모형에서 상호작용이 없고, 마지막으로 영모형(null model)은 축소모형에서 문항 점수가 없는, 즉 대응변수만 고려한다. 따라서, 일방적인 차별기능 (uniform DIF) 문항을 검사하기 위해서 축소모형과 영모형의 우도함수를 포함하고, 비일방적인 차별기

능(nonuniform DIF) 문항을 평가하기 위해서는 축소모형과 완전모형의 우도함수를 포함한다. 일단 일방적 혹은 비일방적 DIF가 유의도 검증에서 밝혀졌다면, 적어도 문항 반응의 한 수준(e.g., 본 연구의 문항들은 세 가지 문항 점수를 가지므로 U=0,1, or 2) 에서 어떤 소속집단의 확률은 문항점수와 대응변수 점수가 주어졌을 때, 오로지 대응변수 점수만으로 예측했던 것과는 유의하게 달라진다라는 결론을 내릴 수 있다.

이런 유의도 검증을 실시한 후, DIF의 실제적 심각성(actual severity)과 DIF의 방향을 알아보기 위한 하나의 사후검사로서 이 로지스틱 판별분석에서는 그래프 기법을 이용한다. 즉, 유의한 DIF를 보였던 문항들을 이용하여 각 문항 점수 수준 (U= 0, 1, 2)에서 추정된 판별함수 주변에 95% Scheffe 유형의 신뢰구간을 형성하고, 그것을 영모델을 위한 판별함수와 비교하는 것이다(Miller & Spray, 1993). 영모델이라는 것은 문항 반응이 제외된 모델로서, 이는 어떤 문항에서든 동일하여, 이는 no-DIF 회귀(no-DIF regression)로서의 역할을 한다. 만약 이 신뢰구간이 대부분의 대응변수 점수에서 영모델을 포함한다면, 그 문항에서 DIF는 실제적으로 심각한 정도는 아니라고 간주한다.

2. SIBTEST (Simultaneous Item Bias Test)

Shealy와 Stout (1993)는 DIF를 한 검사의 문항들에서 일원성(unidimensionality)이 위배되면 나타나는 것으로 개념화 시켜면서, Simultanous Item Bias Test (SIBTEST)를 개발하였다. 이 검사는 처음에는 오로지 이분문항만을 다룰 수 있는 검사로 출발하였으나, 현재 비일방적인 DIF ("Crossing DIF" 라고도 부름)를 추출할 수 있고 (Li & Stout, 1993), 다분 문항을 다룰 수 있으며 (Chang, Mazzeo, & Roussos, 1993), 두 가지 차원의 검사 자료에서 DIF를 추출할 수 있는, 즉 2개의 대응변수를 다룰 수 있는 (Stout, Li, Nandakumar, & Bolt, 1997) 검사들을 개발하고 있다. 이 SIBTEST의 이론적인 구조하에서, 예컨데, 한 성취도 검사는 하나의 목표 능력 (target ability), θ 으로 구성되어 있지만, 문항반응들이 하나 혹은 그이상의 잡음 결정인자 (nuisance determinants), η 에 의하여 결정될 때, DIF에 대한 잠재성이 나타난다고 정의하고 있다. 그러나, 한 검사의 문항들에서 다차원성 (multidimensionality)이 존재한다고 해서 그것이 자동적으로 DIF를 일으키는 것은 아니다.

이 SIBTEST는 검사 문항들을 두 개의 하위 검사, 즉 대응 하위검사 (matching

subtest)와 연구되어지는 하위검사 (studied subtest) 혹은 단일 문항(a studied item)으로 분리를 한다. 각 준거집단과 초점집단에 속하는 피험자들은 그들의 대응 하위검사 점수에 따라서 J 개의 그룹으로 나뉘어 지며, 대응 하위검사에서 같은 점수를 가지고 있는 서로 다른 두 개의 그룹 (준거 혹은 초점 그룹)에 속한 피험자들을 각 문항(혹은 studied subtest)에 대한 그들의 수행으로 비교하는 것이다. 이 검사에서는 다음과 같은 영가설과 대안가설을 평가하게 된다.

Ho: $\beta_U = 0$, Ha: $\beta_U \neq 0$

β_U 라는 것은 하나의 문항이 연구되어지는 경우 일방적 DIF의 양을 나타내는 모수로서 다음과 같은 공식에 의하여 계산되어진다.

$$\beta_U = \sum_{j=1}^{J} p_j (Y^*_{Rj} - Y^*_{Fj}) \qquad (2)$$

여기에서 P_j 는 하위그룹 j 에서, 초점 그룹 구성원의 비율을 나타내며, Y^*_{Rj} 와 Y^*_{Fj} 은 준거집단과 초점집단에 있는 피험자들의 연구되어지는(studied) 문항 혹은 문항들의 조정된 평균들이다 (각 하위집단 j에서). 즉, SIBTEST는 고전진점수이론에 근거하여 잠재변인 (e.g., 능력, 혹은 외향성)을 추정 (즉, 대응변수의 진점수)하고, 그 추정치 위에 문항 수행을 회귀시킨다. 각 대응변수 점수 수준에 따라서, 준거집단과 초점집단의 문항-진점수 회귀선의 차이에 대한 평균을 낸 것이다. 여기에서 유의미한 양수의 β_U 값은 준거집단에 유리하게 기능하는 문항이고, 유의미한 음수의 β_U 값은 초점집단에 유리하게 기능하는 문항이다.

특히, 대응변수의 선택은 차별문항기능 분석에서 중요하며, 이 SIBTEST의 이론적 구조하에서는 몇 가지 방법으로 대응변수 점수를 선택할 수 있다(Shealy & Stout, 1993). 첫째, 문항내용이나 고전 검사 통계치들을 검토하여 전문가적인 판단에 근거하여 대응변수를 구성할 문항을 선정할 수 있다. 그러나, 이는 흔히 영향(impact)과 DIF를 혼돈할 가능성이 있으므로 주의가 요구된다. 두 번째로는 자동적인(automatic) DIF 분석을 실시하여, 자동적으로 각 문항을 DIF를 위하여 검토하고, 그 문항을 제외한 나머지 문항들을 대응변수로서 사용하는 것이다. 본 연구에서는 이 자동적인 DIF분석을 사용하였다.

3. 두 추출 기법간의 비교

Potenza와 Dorans (1995)는 다분 문항을 위한 차별문항기법들을 두 가지 차원에 의하여 분류하였다. 첫째로, 대응변수로 사용되는 특성 추정치(trait estimate)의 성격, 즉 그것이 관찰된 점수인지 잠재 변인인지에 따라서 관찰 점수(observed score) 접근법과 잠재 변수 (latent variable) 접근법으로 분류하였다. 두 번째로는 각 특성의 수준에서 문항 수행(item performance)이 결정되는 방식이 수학적 함수에 의해 추정하는 것이면, 모수적(parametric) 기법, 그렇지 않으면 비모수적(nonparametric) 접근으로 분류하였다. 이 분류에 따르면, LDFA는 관찰된 점수를 대응변수로 사용하는 모수적 기법으로 분류되고, SIBTEST는 잠재적 특성의 추정치를 대응변수로 사용하는 비모수적 기법이다.

다른 DIF 기법들과는 달리, LDFA는 처음부터 다분문항의 사용을 위하여 개발되어진 것으로 (Miller & Spray, 1993), 표본의 크기만 적절히 크다면, 다분문항에서 DIF를 추출하는데 상당한 검증력(power)을 지니고 있다. 또한 이 기법은 두 가지 유형의 DIF 모두를 추출할 수 있고, 다변량 대응변수를 다룰 수 있는 신축성이 있으며, 시각적으로 DIF의 정도와 방향을 알 수 있도록 해주는 사후 기법, 즉 그래프 기법을 제공한다는 장점이 있다. 그러나, 이 기법은 적절하게 모수를 추정하기 위하여서 상당히 큰 표본을 필요로 한다는 단점이 있다(e.g., $n > 1500$). 반면, SIBTEST는 준거집단과 초점집단의 능력 분포 (혹은 특성 분포)가 동질적이지 않은 경우 상당히 항내적(robust)이라는 장점이 있다(Shealy & Stout, 1993). 하지만, 이 SIBTEST는 오직 일방적인 DIF만을 추출할 수 있다는 한계점이 있다. 현재 비일방적 DIF를 추출할 수 있는 SIBTEST는 개발(Chang et al, 1993) 중 이어서, 본 연구에서는 오로지 일방적 DIF만을 추출할 수 있는 PSIBTEST를 사용하였다.

두 개의 DIF 기법을 사용하였기 때문에 예상되는 4가지 종류의 결과가 있다. 우선 두 기법으로부터 나온 결과가 일치를 하는 경우이다. 즉, 두 기법 모두 DIF를 발견하는 경우 또는 모두 DIF를 발견하지 못하는 경우로서, 이 경우 DIF의 존재 여부에 대한 강한 증거를 제공하는 결과라고 볼 수 있다. 세 번째와 네 번째는 두 기법 중 오로지 한 기법만이 DIF를 발견하는 경우들이다. 만약 LDFA만이 DIF를 추출했다면, 그 DIF의 유형이 비일방적이기 때문일 수도 있지만, 두 그룹간의 특성변인의 분포 차이 때문에 생기는 오차로 인해 LDFA가 잘못 기능 했을 수도 있다. 반면, SIBTEST

만이 DIF를 발견한 경우라면, 표본크기가 충분히 크지 않아서 LDFA가 모수를 적절히 추정할 수 없기 때문에 생긴 오차일 수도 있다. 그러나 만약 충분히 큰 표본을 사용했는데도, LDFA가 DIF를 발견하지 못하였다면, 우리는 해석에 있어서 극도의 주의가 필요할 것이다.

Ⅲ. 연구 방법

1. 연구대상

이 연구에서는 문화와 언어가 다른 두 개의 그룹, 즉 한국 대학생 538명, 그리고 미국대학생 844명을 피험자로 사용하였다. 2000년 1월부터 5월에 걸쳐, 한국 표본은 한국의 두 개의 사립대학교로부터 얻어졌고, 미국 표본 중 431명은 University of Illinois에서, 413명은 Institute for Personality and Ability Testing (IPAT)으로부터 얻어졌다. 이 연구에서 피험자를 대학생으로 국한시킨 이유는 비교문화연구에서 가장 중요한 문제라고 볼 수 있는 비교 가능한(comparable) 표본의 선정을 위하여 가능한 한 같은 직업 군에 속하고, 비교적 비슷한 교육 수준을 가지고 있는 피험자를 선정하기 위한 것이었다. 또한 대학생 표본은 검사실시가 용이한 표본이며, 정상 성격 규준(norm)에 비추어 볼 때, 비교적 정상 범위내의 성격 프로파일을 보이는 경향이 있다. 이 두 표본의 배경변인에 관한 정보는 <표1>에 요약되어져 있다.

2. 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 널리 사용되어 오고 있지만, 이 검사는 영어판 16PF와 비교해 볼 때 상당 부분 수정되어 있기 때문에(예컨데, 문항 수, 문항반응양식, 문항 내용 등) 문항 단위에서 평형성(equivalence at item level)을 알아보고자 하는 본 연구의 목적에는 적절치 않았다. 따라서, 본 연구에서는 성격요인검사 (염태호 & 김정규, 1990) 대신, Shaughnessy와 Kang(1998)이 그들의 영재연구를 위하여 사용하였던 한글판 16PF 검사를 부분 수정하여서 사용하였다. 16PF 검사는 세 가지

수준의 강제 선택 문항 양식(a three level forced choice response format)을 사용하며, 추리(reasoning) 척도를 제외한 나머지 16개 척도에서 "그렇다", "잘 모르겠다 혹은 둘 다 그렇지 않다 (중립적 반응)", 그리고 "그렇지 않다"라는 방식의 반응양식을 가지고 있다. 이 검사는 개인 혹은 집단 검사 모두가 가능하고, 검사 실시 시간은 약 45분 정도 소요된다. 이 검사는 16개의 주요 요인(primary factors)들과 추가적으로 Impression Management (IM) 척도라는 자신의 인상을 조절하려는 수준을 측정하는 일종의 허위 척도를 포함하고 있다. 이 16개의 주요 요인들은 외향성(extraversion), 불안감(anxiety), 강인함(tough-mindedness), 독립성(independence)과 자기통제(self-control), 즉 5가지의 이차 요인들로 (secondary or global factors) 다시 묶여질 수가 있다. 이 이차 요인들 가운데, 본 연구에서는 51개의 문항들로 구성된 외향성 요인만을 살펴볼 것이며, <표2>에 외향성 요인을 구성하는 하위척도에 대한 설명이 요약되어져 있다.

3. 연구 분석 절차

가. 한글판 16PF 검사

Shaughnessy와 Kang (1998)이 마련한 한글판 16PF 검사를 기초로 하여, 본 연구에서는 검사 번역가로서의 기준에 적절히 부합되는 두 명의 박사과정 학생들을 선정하여 이 한글판 16PF 검사의 번역 수준을 평가하도록 하였다. 여러 번의 편집 회의(editorial review)를 거듭하여, 이 두 명의 번역가로부터 최종 한글판 16PF 검사 문항들이 얻어 졌다.

나. 문항의 채점

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

다. 일원 차원성 (unidimensionality) 평가

차별문항 기능 분석에 앞서서 자료의 차원성을 평가해 보는 절차는 항상 필요하다. 즉, 검사를 구성하고 있는 문항들이 그 검사가 측정하고자 하는 예컨데, 능력, 성격, 혹은 성격과 같은 하나의 구성요인(one trait)을 재고 있는지를 반드시 평가해 봐야 한다. 만약 문항들이 그 검사가 재고자 하는 구성요인 이외의 다른 것을 재고 있다고

한다면 그것은 문항편파를 일으키는 하나의 원인이 될 수 있기 때문이다. 각 인종 집단 별로 각 척도의 차원성을 평가하기 위해서 본 연구에서는 주성분 분석(Principal Component Analysis)과 확인적 요인분석(Confirmatory Factor Analysis)을 실시하였다. 각 인종그룹별로, 외향성의 다섯 가지의 하위척도들은 개별적으로 평가되었으며, 먼저 PRELIS 2 (Jöreskog, & Sörbom, 1986)을 이용하여 주성분 분석이 실시되었다. Reckase(1979)에 따르면, 일차원성을 가지고 있는 검사는 하나의 주된 요인들로 구성되어 있고, 그 주된 요인의 고유값(eigenvalue)이 나머지 요인들의 고유값들보다 훨씬 커야 한다고 하면서, 적어도 그 주된 요인이 전체 분산의 20%이상을 설명해야 한다고 주장하였다. 더불어서 PRELIS 2를 이용하여 각 인종그룹별로 다분상관계수(polychoric correlation matrix)를 구하고, LISREL 8 (Jöreskog, & Sörbom, 1993)을 이용하여 확인적 요인분석(CFA)을 실시하였다.

라. 차별 기능 문항 분석

문항 수준에서의 평형성(equivalence at item level)을 평가하기 위하여 두 가지의 차별 기능 문항 기법이 사용되었다.

(1) 로지스틱 판별 분석 (LDFA)

통계 프로그램 SAS 8.1의 PROC LOGISTIC 절차를 DESCENDING 옵션과 함께 사용하여 로지스틱 판별 분석(Miller & Spray, 1993)을 실시하였다. 준거(reference)집단과 초점(focal) 집단이 각각 "1" 그리고 "0"으로 코딩되어졌기 때문에, 판별계수가 양수이면 준거집단 (미국집단)이 유리하게 그리고 음수이면 초점 집단(한국집단)이 유리하게 기능하는 문항임을 의미한다. 사후 기법으로서 DIF의 종류, 실용적인 중요성, 그리고 DIF의 근원(locus)을 알아보기 위하여 그래프가 이용되어졌다.

(2) PSIBTEST

다분 문항을 위한 SIBTEST(Shealy & Stout, 1993), 즉 PSIBTEST에서는 두 집단 즉, 준거집단 또는 초점 집단 모두에 대립하여 DIF를 나타내는 문항들을 구별하기 위하여 "e" 옵션과 함께 자동적인(automatic) DIF 절차를 이용하였다.

(가) 차별 기능 문항의 양 (Amount of DIF)

β는 PSIBTEST에서 DIF의 양을 나타내는 계수로서 Roussos와 Stout(1996)의 이분문항(dichotomous items)을 위한 지침에 따르면 영가설이 기각되었을 때, 만약 $|\beta|$

< 0.059 이면, DIF는 무시할 만큼 작은 정도(negligible) 혹은 "A-수준의 DIF"; 만약 0.059 ≤ |β| ≤ 0.088 이면, DIF는 중간 정도 수준(moderate) 혹은 "B-수준의 DIF", 그리고 만약 |β| ≥ 0.088 이면 DIF는 큰 정도(large) 혹은 "C-수준 DIF" 라고 한다. 이 지침을 다분문항에 적절하게 변환시키기 위하여서는, β의 계수를 각 문항의 점수 범주의 수로 나누어 주면 된다. 예컨데, 16PF 검사는 세 개의 반응 옵션을 가지고 있고, DIF의 양을 알아보기 위하여 β의 값을 "2"로 나누어 주면 된다. 본 연구에서는 C 수준의 DIF만을 유의미한 수준의 DIF로 간주하였고, LDFA와 SIBTEST, 두 기법 모두에서 공통적으로 추출된 DIF는 모두 C 수준에 해당하는 차별기능으로 밝혀졌다.

(나) 순환적인 정수과정 (Iterative Purification Process)

이 순환적인 정수기법은 주로 내적인 기준(internal criterion)만이 대응변수(matching variable)로 사용되어질 경우 효과적인 것으로 알려져 있다. 자동적인 DIF 분석에 따라서, 거의 모든 문항들이 편파 되어졌다고 밝혀졌다면 이 순환적인 정수기법을 사용하여 최종적으로 타당한 문항들로만 구성된 일군의 문항들이 남을 때 까지 DIF 분석을 반복하도록 하는 것이다. 즉, 가장 큰 SIB 통계치와 가장 작은 p 값을 가지고 있는 문항을 먼저 대응변수에서 제외시키고 DIF 분석을 반복한다. 이런 계속적인 반복을 통하여 마침내 타당한 대응변수를 구성할 수 있는 문항들을 얻을 것이며, 이 최종적인 대응변수를 이용하여 DIF 분석을 하는 것이다.

Ⅳ. 연구 결과

1. 기술 통계치(Descriptive Statistics)

<표 3>에는 외향성 하위척도의 점수에 대한 평균과 표준편차에 대한 정보가, <표 4>에는 척도별 신뢰도 계수가 각 인종그룹별로, 그리고 다시 각 그룹내의 성별로 요약되어져 있다. 미국표본에서는 .78의 중앙치(median)를 가지고, .69-.87의 내적 일관성 계수 범위를 보였고, 한국표본에서는 .62의 중앙치와 함께 .54-.87의 범위를 가지고 있었다. 미국표본의 신뢰도 계수는 16PF 검사 매뉴얼(Russell & Karol, 1994)에 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.

2. 일원 차원성 (unidimensionality)

기존의 선행연구 (Ellis & Mead, 1998; Flanagan, Raju, & Hayhood, 1998)에서 16PF의 주요 척도(primary scales)들은 높은 정도의 일차원성을 보여주었다. <표 4>에서 나타난 비교적 높은 신뢰도 계수 역시 부분적으로 일차원성의 증거를 보여주는데, <표 5> 의 주성분 분석의 결과 역시 각 하위척도의 일차원성을 지지해 주고 있다. A, F, N 척도에서 그룹간 설명된 분산의 양을 비교해 보면 한국표본에서 비교적 작은 설명량을 보이지만, Reckase(1979)의 제안처럼, 첫 번째 요인에 의해서 설명된 분산이 5개의 모든 척도에서 20% 이상으로 나타났기 때문에, 우리는 각 하위척도는 두 개의 그룹 모두에서 일차원성의 가정을 만족시킨다고 할 수 있다. 추가적으로 확인적 요인분석의 결과, 5개의 요인으로 구성된 외향성 모델이 두 그룹 모두에서 만족스런 적합도를 보여줌으로써 일차원성 가정을 지지해 주고 있다.즉, 미국그룹에서는 카이자승 = 333.05 (df=80), GFI = .95, RMSEA = .061, RMR = .041 이었고, 한국 그룹에서는 카이자승 = 195.47 (df=80), GFI = .95, RMSEA = .053, RMR = .042의 적합도 지수를 보였다.

3. 인종 그룹에 따른 차별 기능 탐색

1) 차별 기능 문항의 추출

각 하위척도의 점수를 단일 대응변수로써 사용하였고, 연구되어지는 문항(studied item)의 점수는 이 대응변수의 점수에서 제외시켰다. 유의도 수준은 .01을 사용하였고, 얻어진 p 값이 .01 보다 경우, 유의미한 DIF로 간주하였다. LDFA와 PSIBTEST에서 얻어진 결과는 <표 6>에 요약되어져 있다.

먼저, A척도에서는 LDFA는 PSIBTEST가 추출한 5개의 DIF 문항을 모두 포함하는 8개의 문항을 비일방적 DIF로 확인하였다. F척도에서는 2개의 비일방적 DIF 문항과 1개의 일방적 DIF 문항이 LDFA에 의해서 추출되었으며, 4개의 문항이 PSIBTEST에 의해서 차별 기능문항으로 발견되어졌다. 이 F 척도에서는 오로지 두 개의 문항 (68번과 100번 문항)만이 두 추출방법 모두에서 DIF로서 밝혀졌다. H척도를 살펴보면 73번 문항이 두 방법 모두에서 일방적 DIF 문항으로 추출되었고, 105번과 107번 문항은 오로지 PSIBTEST에서만 DIF 문항으로 밝혀졌다. N 척도에서는 143번 문항을 제외한 나머지 9문항 모두가 비일방적 DIF 문항으로 LDFA에 의해서 추출되어졌고, 반면, 6개의 문항이 PSIBTEST에 의해서 차별기능문항으로 판별되었다. Q2 척도를 보면,

121번, 123번, 그리고 156번의 세 개 문항이 두 추출방법 모두에서 일방적 DIF 문항으로 밝혀졌으며, 152번 문항은 PSIBTEST에 의해서만 DIF 문항으로 밝혀졌다.

2) 차별기능문항 추출방법간 비교

두 차별기능 문항 추출방법들로부터 나온 결과를 비교해 보기 위하여, <표7>에 각 방법에 의해서 밝혀진 DIF와 non-DIF 문항의 수에 대한 빈도를 요약해 놓았다. 먼저, 차별기능문항을 추출하는데 있어서 PSBITEST와 LDFA간의 일치(agreement)의 정도를 알아보기 위하여 Cohen's *kappa* (Agresti, 1996)를 사용하였다. 이 *kappa*의 범위는 우연수준에서의 일치도,"0"로부터 완전한 일치도 "1" 까지로서 LDFA와 PSIBTEST의 일치도는 .42수준으로 중간 정도 수준의 일치도를 나타냈다. 더불어서 *McNemar*의 검사(Agresti, 1996)를 이용하여 LDFA와 PSIBTEST가 발견한 차별기능문항의 수에는 유의미한 차이가 있는지 검증해 보았다. 이 결과, 검증 통계치 z 는 .26으로서 .05의 유의도 수준에서 유의미하지 않음을 나타냈다. 즉, LDFA와 PSIBTEST는 차별기능문항을 추출하는데 있어서 통계적으로 유의한 차이는 보이지 않았다.

4. 대응 변수의 선택

1) 다변량 대응 변수 기법 (Multivariate Matching Technique)

번역 DIF를 찾아내려는 연구에서 사용될 수 있는 최상의 대응변수를 모색하려는 시도의 일환으로, 세 가지의 종류의 내적 준거(internal criterion)를 평가해 보았다: a) 각 척도의 점수, b) 각 척도의 점수와 외향성 점수, c) 5개의 개별적인 척도의 점수들. <표 8>에 나타난 각 대응변수간의 상관계수를 살펴보면, 서로 간에 유의미하게 상관되어져 있음을 알 수 있다 ($p < .01$). 로지스틱 판별 분석의 결과에 따르면 각 대응변수의 조합에서 나타난 차별 기능 문항의 수는 상당히 일관적인 것으로 나타났다 (표 9 참조). 이는 여러 개의 대응 변수를 동시에 이용하는 것이 하나의 척도 점수만을 사용하는 것 보다 피험자들을 더욱 정확하게 대응시킬 수 있다는 선행 연구의 결과(Clauser, Nungester, & Swaminathan, 1996; Mazor, Kanjee, & Clauser, 1995)와는 일치하지 않았다.

2) 순환적인 정화 기법 (Iterative Purification Technique)

대응(matching)의 정확성을 높이려는 두 번 째 시도로서, PSIBTEST를 순환적

(iteratively)으로 사용하여 대응 변수를 정화(purify)시켜 보았다. 첫 번째 시행에서 a) A 척도에서 5개의 문항, b) F 척도에서는 4개 문항, c) H 척도에서 3개 문항, d) N 척도에서 6개 문항, e) Q2 척도에서 5개 문항이 DIF를 나타내었다. 두 번째 시행을 위하여, 가장 큰 SIB 통계치를 가지고 있는 문항을 각 대응변수에서 제외시키고 다시 PSIBTEST를 반복하였다. 이런 과정은 모든 문항이 DIF를 나타내거나 너무나 적은 non-DIF 문항이 남을 때까지 반복되었다. 세 번째 시행 이후부터는 반 이상의 문항들이 모두 DIF를 나타내었던, 즉 확산적인 DIF(pervasive DIF) (Ellis & Mead, 1998)를 가지고 있는 A, N, 그리고 Q2 척도에서는 이런 정화과정은 제대로 작동할 수가 없었다. 반면, F와 H 척도에서는 각각 4번과 3번의 시행을 반복하였고, 결과적으로 6개의 F척도 문항과 7개의 H척도 문항이 정화된 대응변수로서 남게되었다. 이 정화된 대응변수만을 이용하여 다시 PSIBTEST를 실행하였으나, 첫 번째 시행과정에서 발견한 동일한 문항들에서 DIF를 발견하였다. 즉, 이 연구에서는 정화된 대응변수의 사용이 차별 기능 문항을 추출하는 데에 어떠한 영향도 미치지 못함을 알 수 있었다.

5. 그래프 기법을 이용한 사후 검증

두 개의 차별기능문항기법 모두에서 DIF를 나타낸 16개의 문항에 대하여 사후절차로서 DIF의 실제적 중요성과 더불어서 DIF의 근원(locus)를 알아보기 위하여 완전모형, 그 주변의 95% 신뢰구간, 그리고 영모형을 그래프로 나타냈다(그림1 참조). 16개의 그래프를 살펴보면, 영모형이 완전모형의 신뢰구간으로부터 거의 모든 대응변수 점수(X) 범위에서 이탈되거나 아주 적은 점수 범위에서만 포함되어졌다. 이는 16개의 문항에서 나타난 DIF는 의미있는 것으로 간주될 수 있다는 정당성을 제공해주는 것이다 (Miller & Spray, 1993). 또한 16개의 문항 중 4개의 문항(159, 98, 47, & 84)은 극단 점수, 즉 U = 0 (반대) 혹은 U = 2(긍정)에서 의미있는 DIF를 나타내고, 중간점수, 즉 U = 1 (중립적 의견)에서는 DIF를 보이지 않았다. 그러나, 나머지 12개 문항들에서는 거의 모든 문항점수, 즉 U = 0, 1, 2 에서 의미있는 DIF를 보였다. 이 결과는 중립적인 반응을 나타내는 중간의 문항 점수 (U=1)에서 조차도 인종별 DIF를 나타낼 수 있다는 단서를 제공하는 것으로써, 만약 이 문항들을 이분화시켰다면 우리들에게 또 다른 결과를 제공하고, 이에 대한 정보를 손실할 지도 모른다는 가능성을 제시한다.

V. 논 의

본 연구에서는 외국의 교육 및 심리검사를 도입하여 비교연구에 사용하기에 앞서서, 검사 점수의 "평형성" 확립이 무엇보다도 중요하다는 관점에서 출발하였으며, 이런 평형성을 평가하는 것은 질적인 방법 뿐 아니라, 통계적인 방법도 함께 수반되어야 한다는 점을 강조하였다. 특별히, 차별기능문항 기법을 이용하여, 서로 다른 언어와 문화 그룹에서 문항 수준에서의 평형성이 유지되고 있는지를 살펴보았다. 우선 DIF의 실제적 중요성을 고려하여, 두 개의 DIF 기법 모두에서 차별기능을 나타내는 것으로 밝혀진 16개의 문항만(총 51개 문항 중에서)을 검토하였다. 연구 결과를 요약하자면, 외향성 척도에서 총 31%의 문항이 차별기능을 나타내었고, 각 척도별로 살펴보자면, A척도의 45% 문항, F척도의 20%문항, H척도의 10%문항, N척도의 50% 문항, 그리고 Q2척도의 30% 문항이 DIF를 나타내었다. 이 결과는 중국어 버전과 스페인어 버전(Zhang & Yan, 2000)의 16PF(Ellis & Mead, 1998)를 다루었던 두 선행연구 결과와 어느 정도 일치되는 것으로 이 두 연구에서도 H척도에서 가장 적은 수의 DIF 문항을 발견하였고, N척도에서 가장 많은 수의 DIF 문항을 발견하였다. 따라서, H척도는 비교적 비교문화장면에서 사용하기 쉬운 문항편파가 없는 (DIF-free) 문항들로 구성되어 있다고 볼 수 있지만, 다른 척도들을 이용하여 서로 다른 언어 및 문화 그룹간 비교를 할 때는 상당한 주의가 필요할 것이다. 더불어서, 차별기능문항의 수와 각 척도의 일차원성 정도와 밀접한 관련이 있음을 발견할 수 있었다. 즉, 신뢰도 계수가 낮거나, 주성분 분석에서 첫 번째 요인이 설명한 분산의 양이 한국 그룹에서 상대적으로 적은 척도들, 즉 A와 N 척도에서 더 많은 DIF를 나타내는 경향이 있었다.

본 연구에서는 서로 다른 접근법을 가지고 있는 LDFA와 SIBTEST와 같은 두 개의 차별기능기법을 상호보완적인 목적으로 사용하였으며, 언어와 문화가 서로 다른 두 그룹의 구성원들을 보다 정확하게 대응시키려는 노력으로 두 가지 기법이 시도되었다. 단일 그룹을 대상으로 학업성취도 검사와 같은 인지적 검사를 다루었던 선행연구(예: 성별에 따른 DIF)에서는 성공적으로 적용되었던 다변량 대응기법(Clauser, Nungester, Mazor, & Ripkey, 1996)과 순환적인 정화기법을 본 연구에 적용해 보았지만 별다른 효과를 가져오지 못하였다.

이 연구의 결과를 기초로 하여 앞으로 진행될 수 있는 몇 가지 연구에 대한 제안을

해 보고자 한다. 첫째, 언어와 문화가 다른 그룹간에 점수의 평형성을 검토할 수 있는 통계적인 방법들의 개발과 사용에 초점을 두는 연구가 계속되어야 한다. 본 연구는 서로 다른 그룹들을 보다 정확하게 대응시킬 수 있는 방안으로 다변량 대응기법이나, 순환적인 정화기법을 적용해 보았으며, 비교문화 장면에서 외국검사를 번안할 때 사용할 수 있는 가장 적절한 대응변수에 대한 더 많은 후속 연구가 필요할 것이다. 또한 문항 수준에서의 평형성 뿐 아니라, 실제적인 개인에 대한 결정은 문항 수준이 아니라 하위척도 수준에서 이루어 진다는 점을 감안한다면(Ellis & Mead, 1998), 일정한 하위척도 수준(at scale level)에서의 평형성을 검토하는 것도 의미 있는 일이 될 것이다. 이를 위하여, 일정한 그룹의 문항이나 척도 수준에서의 차별기능을 밝혀 낼 수 있는 SIBTEST나, Raju가 제작한 differential item and test functioning (DFIT) 기법(Raju, van der Linden, & Fleer, 1992)들이 유용할 것이다. 또한, 한 척도의 50% 이상의 문항들이 차별기능을 나타내는 경우는 상호보완적인 DIF (compensatory DIF) 기법의 활용도 생각해 볼 수 있을 것이다. 최근에는 다그룹 확인적 요인분석의 확장적인 형태인 MACS(Analyses of the means and covariance structures) 역시 문화적인 편파를 다루는 대안으로 소개되고 있다(Chueng & Rensvold, 2000; Little, 2000). 이러한 통계적인 방법들을 적용하고 개선시킴으로써, 우리는 비교문화연구에서 사용할 수 있는 더욱 타당하고 신뢰로운 검사들을 개발할 수 있는 기회를 확장시킬 수 있을 것이다.

둘째, 앞으로의 연구들은 DIF를 나타내는 문항들의 원인을 규명하기 위한 노력이 필요할 것이다. 번역된 문항들에서 DIF의 원인을 찾아내는 것은 의외로 일반적인 DIF의 연구에서 원인을 밝혀내는 것보다 더 명확할지도 모른다. 가장 큰 원인은 번역과 각 문항내용의 문화적 적절성과 많은 관련을 가질 것이기 때문이다.

마지막으로, 한 검사의 비교문화적인 평형성(cross-cultural equivalence)의 확립은 단일한 하나의 방법에 의해서 이루어 질 수 없는 것이고, 다양한 종류의 평형성을 다룰 수 있는 다양한 방법이 사용되어져야 하는 연속적인 과정이라는 것이다. 우선, 검사를 번역하고 번안하는 과정에서 상당한 시간과 노력을 투자하여야 하며, 일단 번안된 검사를 비교문화연구에 사용하기에 앞서서 통계적인 방법과 질적인 방법을 이용하여 다각적으로 검사의 평형성을 검토해야 할 것이다.

<참 고 문 헌>

성태제 (1994). 1994학년도 제1차 대학수학능력시험의 성별에 따른 차별기능문항 추출. ***교육평가연구. 7***(2), 87-101.

송미영 (2001). 다분 차별기능 문항의 추출방법들을 이용한 수행평가 과제의 차별기능 탐색. ***교육평가연구, 14***(1), 81-101.

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

Agresti, A. (1996) ***An introduction to categorical data analysis.*** NY : John Wiley & Sons, Inc.

Bontempo, R. (1993). Translation fidelity of psychological scales: An item response theory analysis of an Individualism-Collectivism Scale. ***Journal of Cross-Cultural Psychology, 24***(2), 149-166.

Brown, P. J. (1996, April). *Using differential item functioning analysis to determine differential interpretations of survey questions.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. ***Applied Psychological Measurement, 19***(4), 309-321.

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.

Chang, H., Mazzeo, J., & Roussos, L. (1993, April). *Extension of Shealy-Stouts DIF procedure to polytomous scored items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Cheung, G. W., & Rensvold, R. G. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. ***Journal of Cross-Cultural Psychology, 31***(2), 187-212.

Church, A. T., & Lonner, W. T. (1998). The cross-cultural perspective in the study of personality : Rationale and current research. ***Journal of Cross-Cultural Psychology, 29***(1), 32-62.

Clauser, B. B., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. ***Journal of Educational Measurement, 33***(4), 453-464.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. ***Journal of Educational Measurement, 33***(2), 202-214.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relatiosnwith 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.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). ***Journal of Cross-Cultural Psychology, 24***(2), 133-148.

Ellis, B. B., Minsel, B., & Becker, P. (1989). Evaluation of atttidue survey tranlsations: An investigation using item response theory. ***International Journal of Psychology, 24,*** 665-684.

Flanagan, W., Raju, N. S., & Haygood, J. M. (1998, August). *Impression management, measurement equivalence, and personality factors: Can IRT be used to determine the impact of faking.* Paper presented at the 13th

annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

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.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits differential item functioning in the NEO personality inventory. ***Journal of Cross-Cultural Psychology, 28***(2), 192-218.

Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of the Jop Descriptive Index into Hebrew. ***Journal of Applied Psychology, 71***(1), 83-94.

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.

Johanson, G. A., & Johanson, S. N. (1996, April). *Differential Item Functioning in survey research.* Paper presented at the Annual Meeting of the American Educational Research Association. New York, NY.

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

Jöreskog, K., & Sörbom, D. (1993). ***LISREL 8.3: Structural equation modeling with the SIMPLIS command language.*** Chicago : Scientific software international.

Kim, A. & Lim, E-Y. (1999, April). *Comparison of effectiveness among different types of practices in cross-cultural test adaptation of attitude measures.*

Paper presented at Annual Meeting of American Educational Research Association, Montreal, Canada.

Li, H., & Stout, W. (1993, June). ***A new procedure for detection of crossing DIF/bias.*** Paper presented at the annual meeting of American Education Research Association, Atlanta, GA.

Little, T. D. (1997). Mean and Covariance Structures (MACS) analyses of cross-cultural data: Practical and theoretical issues, ***Multivariate Behavioral Research, 32***(1), 53-76.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. ***Journal of Educational Measurement, 32,*** 131-144.

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.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. ***Educational Measurement : Issues and Practice, 19***(3), 5-15.

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.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1992, April). *An IRT-based internal measure of test bias with applications for differential item functioning.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. ***Journal of Educational Statistics, 4***(3), 207-230.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. ***Journal of Educational Measurement, 33,*** 215-230.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? ***Educational and Psychological Measurement,*** *59*(2), 248-269.

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

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. (in press). Using bilinguals to evaluate the comparability of different lanugage versions of a test. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.) ***Adapting educational and psychological tests for cross-cultural assessment.*** Hillsdale, NJ: Lawrence Erlbaum.

Sohn, W. J. (1999, April). *Detection of differential item functioning In attitudinal measurement.* Poster presented at Annual Meeting of National Council on Measurement in Education, Montreal, Canada.

Stout, S., Li, H., Nandakumar, R., & Bolt, D. (1997). MULTISIB : A procedure to investigate DIF when a test is intentionally two-dimensional. ***Applied Psychological Measurement,*** *21*(3), 195-213.

Zhang, Z., & Yan, G. (2000). *Differential Item Functioning in the 16PF Questionnaire: A cross-cultural comparison in items and traits.* Unpublished paper.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? ***Journal of Educational Statistics, 15,*** 185-197.

ABSTRACT

Application of Differential Item Functioning to Test Adaptation

Wonsook Sohn

This paper is concerned with evaluating the fidelity of a non-cognitive test adaptation for use in multiple languages and cultures using two differential item functioning(DIF) techniques: (a) PSIBTEST, and (b) Logistic Discriminant Function Analysis(LDFA). In particular, this study focused on how DIF research can best be extended to the problem of evaluating the equivalence of tests across cultures and languages. The Sixteen Personality Factor (16PF) questionnaire was administered in English to 844 American college students and in Korean to 538 Korean college students. This study attempted to identify the best matching criterion for the translated tests by using both a multivariate matching technique and an iterative purification process. The results generally showed a small number of DIF items on each scale, except for scales A and N where about half of the items showed DIF. The choice of matching variables based on a combination of internal measures appeared to have little effect and the iterative purification method was unsuccessful. Finally, the results were discussed and methodological implications were also presented.

<표 1> 배 경 변 인

	미국 피험자 (n=844)			한국 피험자 (n=538)	
	n	퍼센트		n	퍼센트
성별					
여자	224	26.5		207	38.5
남자	575	68.1		319	59.3
Missing	45	5.3		12	2.2
학년					
1학년	52	6.2	(12.1)[a]	198	36.8
2학년	109	12.9	(25.3)	121	22.5
3학년	56	6.6	(13.0)	160	29.7
4학년	92	10.9	(21.3)	22	4.1
대학원생	41	4.9	(9.5)		
Missing	494	58.5		37	6.9
전공					
사회과학/인문/교육	311	36.9	(76.2)	314	58.4
공학/자연과학	39	4.6	(9.1)	106	19.7
경영/경제	3	0.4	(0.7)	50	9.3
예술/건축	40	4.7	(9.3)	50	9.3
Missing	451	53.4	(4.7)	18	3.3

(주1) IPAT으로부터 얻어진 413명에 대한 학년과 전공에 대한 정보는 얻을 수가 없었음.
(주2) [a]퍼센트는 missing data를 제외하고, 431명에 대한 자료에 근거하고 있음.

<표 2> 외향성(Extraversion) 하위척도

요 인	척 도	문항수	낮은 스텐점수(-)	높은 스텐점수(+)a
A	온정성 (Warmth)	11	냉정한 마음씨 말수가 적고, 수줍어함.	다정한 마음씨 낙천적 남과 잘 협력하는
F	쾌활성 (Liveliness)	10	신중하고 말이 적고, 내성적인 성격	쾌활하고, 사교적, 재치있음. 충동적
H	사회적 대담성 (Social Boldness)	10	소심하고, 일을 도모하기가 힘듬. 외적 위협에 민감.	모험을 즐기고, 얼굴이 두꺼워서 사회적으로 상당히 대담.
A	내밀성 (Privateness)	10	솔직함	눈치 빠르고, 교활함. 세속적인
Q2	자기 의존성 (Self-Reliance)	10	집단의 결정을 따르고, 어떤 집단의 추종자가 되기 쉬움	자급자족적이고, 자신의 판단을 따름.

(주) 스텐(sten)은 "standard ten"의 약자로서 M=5.5, SD=2인 표준점수(standard score)이다.

<표 3> 외향성 하위척도의 원점수 평균과 표준편차

하위척도	미국 표본						한국 표본					
	남자 (n = 224)		여자 (n = 575)		합계 (n = 844)		남자 (n = 200)		여자 (n = 321)		합계 (n = 522)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
A	14.14	4.33	17.25	3.90	16.31	4.26	12.70	3.49	14.42	3.86	13.77	3.81
F	14.33	4.27	14.28	4.33	14.27	4.30	12.39	3.43	11.35	4.27	11.76	4.00
H	10.38	6.16	10.84	6.53	10.76	6.40	9.75	5.85	10.86	6.25	10.45	6.13
N	11.93	5.01	10.02	5.27	10.56	5.26	11.39	3.75	11.40	4.33	11.37	4.14
Q2	8.50	4.90	7.14	5.14	7.59	5.11	7.11	4.17	7.97	4.73	7.62	4.55

<표 4> 외향성 하위척도의 신뢰도 계수[a]

하위척도 (문항수)	미국 표본			한국 표본		
	남자 (n = 224)	여자 (n = 575)	합계 (n = 844)	남자 (n = 200)	여자 (n = 321)	합계 (n = 521)
A (11)	.63	.67	.69	.46	.55	.54
F (10)	.69	.71	.71	.49	.67	.62
H (10)	.85	.88	.87	.85	.88	.87
N (10)	.75	.79	.78	.48	.67	.61
Q2 (10)	.75	.80	.79	.71	.78	.75

(주) [a]Cronbach's Alpha : 내적 일관성 계수

<표 5> 주성분 분석의 결과

	미국 표본 (n = 844)			한국 표본 (n = 521)		
	PC_1	PC_2	PC_3	PC_1	PC_2	PC_3
A	3.81[a](34.6)[b]	1.62(14.7)	0.97(8.8)	2.57(23.4)	1.36(12.4)	1.28(11.7)
F	3.88(38.8)	1.15(11.5)	0.92(9.2)	3.00(30.0)	1.37(13.7)	1.04(10.5)
H	6.08(60.8)	0.96(9.6)	0.59(5.9)	5.96(59.6)	1.00(10.0)	0.64(6.4)
N	4.69(46.9)	1.20(12.0)	0.92(9.2)	2.93(29.3)	1.34(13.4)	1.11(11.2)
Q2	4.76(47.6)	1.02(10.2)	0.98(9.8)	4.29(42.9)	1.06(10.6)	0.85(8.5)

(주) [a] 고유값(Eigenvalues) ; [b] 설명 분산량(% Variance) ; PC = 주성분(Principal Components).

<표 6> 인종 그룹별 차별기능기법의 결과

	DIF 유형	Factor A	Factor F	Factor H	Factor N	Factor Q2
로지스틱 판별분석	일방적		68	73		121, 123, 156
	비일방적	1, 33, 96, 127, 159, 65, 98, 129	100, 37		143번을 제외한 모든 문항	
PSIBTEST	비일방적	1, 33, 96, 159, 98	68, 100, 164, 37	73, 71, 105	47, 143, 148, 15, 84, 117	89, 121, 152, 123, 156

(주) 대응변수는 각 하위척도의 총점이 사용되었다. 단, studied item의 점수는 여기서 제외됨.

<표 7> 두 개의 차별문항기능 기법간의 일관성

		LDFA의 결과		
		Non-DIF 문항 수	DIF 문항 수	총합
PSIBTEST의 결과	Non-DIF 문항 수	20	8	28
	DIF 문항 수	7	16	23
	총합	27	24	51

<표 8> 그룹 별 대응 변수들 간의 상관계수

	Factor A	Factor F	Factor H	Factor N	Factor Q2	EXT
Factor A		.31**	.21**	-.43**	-.42**	.71**
Factor F	.27**		.43**	-.28**	-.46**	.67**
Factor H	.41**	.31**		-.39**	-.32**	.61**
Factor N	-.33**	-.27**	-.43**		.40**	-.74**
Factor Q2	-.34**	-.47**	-.21**	.37**		-.74**
EXT	.69**	.67**	.65**	-.69**	-.71**	

(주1) 미국 표본(n = 423)을 위한 상관계수는 대각선 위에 위치하고, 한국 표본(n = 506)을 위한 상관계수는 대각선 아래에 위치한다.

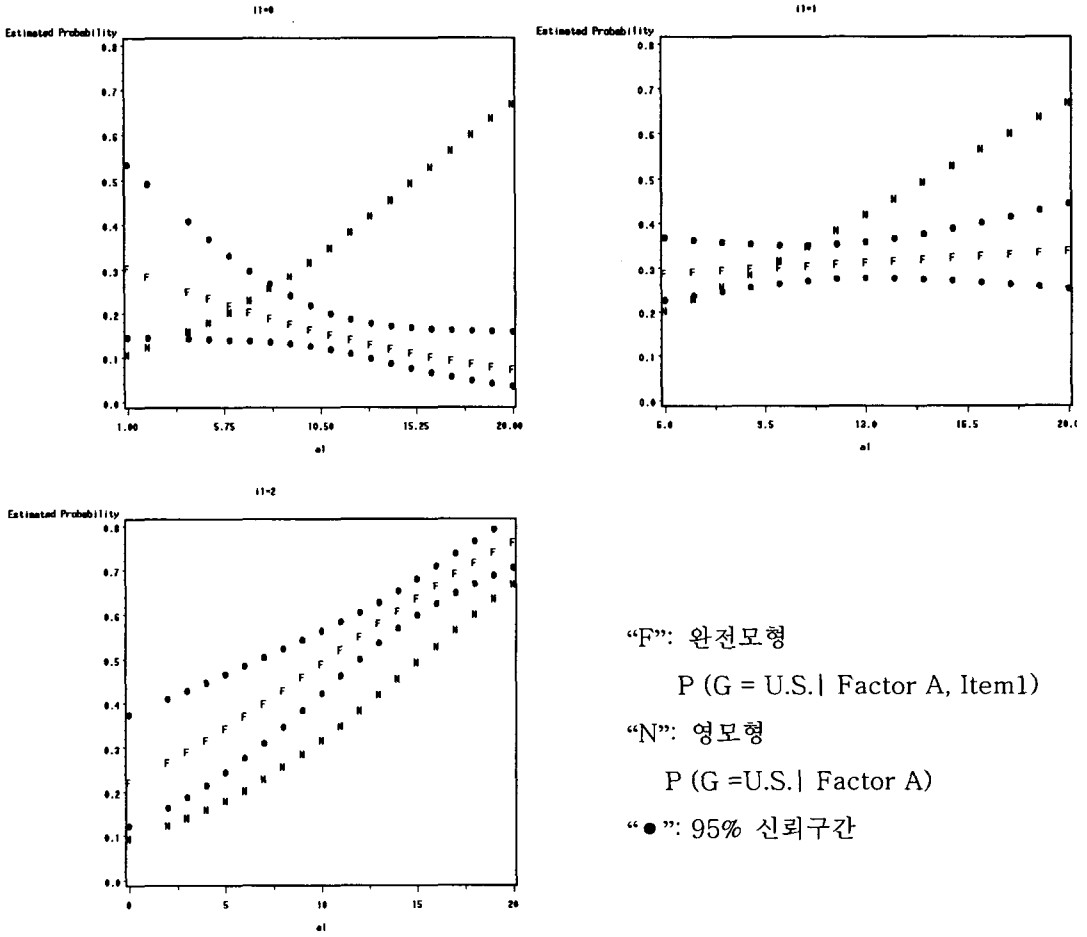
(주2) ** $p < .01$

(주3) EXT = 외향성 점수 = Factors A + F + H + N(reversed) + Q2(reversed).

<표 9> 세 가지 종류의 대응변수에 따른 차별 기능 문항의 수

대응변수	Factor A (11개 문항)	Factor F (10개 문항)	Factor H (10개 문항)	Factor N (10개 문항)	Factor Q2 (10개 문항)
① 각 하위 척도의 점수	8	3	1	9	3
② 1번 & 외향성점수	8	3	1	8	3
③ 5개의 하위척도점수	7	4	1	9	3

(주1) 외향성 점수 = Factors A + F + H + N(reversed) + Q2(reversed) ; $*p < .01$.
(주2) 이 결과는 로지스틱 판별 분석을 이용한 것이다.



[그림 1] 1번 문항의 로지스틱 판별분석 그래프.