勞使關係硏究, 제21권(2010년 8월)

측정동등성의 의미와 검증방법

박 원 우*
양 윤 희**
이 현 정***
최 용 준****
김 문 정**

〈목 차〉

본 연구는 인사조직 분야와 마케팅 분야를 비롯한 경영학, 그리고 넓게는 사회과학 분야에서의 질문지법을 이용한 연구에서 그동안 당연한 가정으로 받아들여졌던 측정동등성의 의미와 영향, 검증하지 않았을 때 발생하는 문제점, 이를 검증하는 여러 가지 방법들을 다루고 있다.

측정동등성이란 시간, 지역, 문화, 집단, 그리고 측정매체 등이 서로 다른 조건 하에서의 현상을 관찰하고 측정하는 작업이 동일한 속성을 측정하는 것을 의미한다. 만약 이것이 검증되지 않은 상태에서 연구가 진행될 경우 연구의 타당도에 심각한 위협을 받게 된다. 본 연구에서는 국내외의 측정동등성에 대한 인식정도를 조사하고자 국내 주요 경영학 관련 학술지인 *경영학연구*, *인사·조직연구*, *인사관리연구*에 게재된 지난 10년간의 논문과 해외 주요 학술지인 *Academy of Management Journal* 및 *Journal of Applied Psychology*에 게재된 지난 5년간의 논문을 대상으로 측정동등성이 검증되어야 할 논문을 선정하고, 각각 측정동등성의 인식 정도와 용어의 사용을 조사해보았다. 그 결과 국내학계의 측정동등성에 대한 인식수준이 국외연구들에 비해 현저히 낮은 것으로 나타났고, 이에 대한 인식수준의 향상과 더불어 연구에서의 측정동등성의 검증을 통해 연구의 타당도를 높이는 노력이 필요하다고 판단하였다.

* 서울대학교 경영대학 교수
** 서울대학교 대학원 경영학과 박사과정
*** 미국 University of Michigan, Survey Methodology Program 석사과정
**** 미국 University of Minnesota, Carlson School of Management 박사과정

측정동등성을 검증하는 방법으로는 대표적으로 세 가지가 자주 거론된다. 구체적으로 확인적 요인분석(confirmatory factor analysis), 문항반응이론(item response theory), 그리고 일반화 가능도 이론(generalizability theory)이 그것이다. 이 각각의 방법에 대한 간략한 의미 및 측정동등성과 연관 지어 어떻게 이러한 방법들을 사용할 수 있는지 그 이론적인 원리와 실증 방법에 대해 알아본다. 이 세 가지 방법에는 각기 장단점이 존재하며 상호 보완하는 역할을 하므로 세 가지 모두 사용하는 것이 바람직하며, 현실적으로 불가능하다면 연구의 설계 유형이나 목적에 따라 적합한 것을 선택하여 사용할 필요가 있다.

I. 서 론

측정이란 자료를 양적으로 분석하기 위하여 일정한 규칙에 따라서 체계적으로 사물이나 현상에 수치를 부여하는 것을 말한다(Stevens, 1951, 1968). 사회과학에서 측정은 매우 중요하면서도 어려운 이슈로 계속 다뤄져 왔는데, 이는 만족도나 소외감 같이 직접 관찰할 수 없는 주관적 인식이나 태도를 실증적으로 조작될 수 있는 구체적인 방법으로 정의하는 것이 어렵고, 그 과정에서 많은 오류가 수반될 수 있기 때문이다. 따라서 연구자들은 진점수(true score)와 오차점수(error score)를 이용한 고전검사이론(classical test theory, CTT)[1]에 근거한 신뢰도와 타당도를 기준으로 삼아 측정의 질을 평가하고 이를 증진시키고자 하는 노력을 경주해왔다. 그러나 기존의 신뢰도와 타당도로는 설명할 수 없는 추가적인 이슈들이 점차 제기되면서 고전검사이론의 한계가 드러나고 있으며, 그런 이슈 중 하나가 측정동등성(measurement equivalence/invariance, ME/I)의 문제이다.

연구자들은 연구를 함에 있어서 그룹간의 비교를 할 수 있도록 우리의 측정도구가 동일한 개념을 다루고 있을 것을 기대한다. 즉, 측정이 그룹간에 '측정동등성'을 만족시킴을 가정하는 것인데(Reise, Widaman, & Pugh, 1993) 이것은 시간, 지역, 문화, 집단, 측정매체 등이 서로 다른 조건 하에서 '현상을 관찰하고 측정할 때, 측정을 위한 조작이 동일한 속성을 측정하고 있는지의 여부'를 말한다(Horn & McArdle, 1992, p.

1) 고전검사이론: 1904년 *American Journal of Psychology*에서 Spearman(1904)이 처음 제시한 교육 및 심리측정의 개념적 모형으로서 검사점수인 관찰점수는 진점수와 오차점수의 합으로 이루어졌다는 것을 기본모형으로 삼는 검사이론

117). 다른 문화권에 있는 응답자들이 동일한 측정도구를 과연 동일한 방식으로 해석하는지, 혹은 상이한 평가자가 동일한 수행 영역에 대해 같은 대상을 평가함에 있어 수행을 일관되게 정의하는가에 대한 질문이 측정동등성 이슈에 대한 예가 될 수 있다. 측정동등성을 고려하지 않은 채 연구를 수행하고 그 결과를 보고한다면, 연구결과가 연구자가 가정한 변수들 간의 관계에서 기인하는 것인지 확신할 수 없기 때문에 결과의 해석에 주의를 기울여야 한다. 예를 들어, Azocar, Arean, Miranda, & Munoz(2001)의 우울 척도 연구에 따르면, 실제 우울함의 정도와 무관하게 스페인계 응답자들은 비 스페인계 응답자보다 "나는 울고싶다"라는 항목에 응답하는 경향이 높았다. 라틴 문화는 우는 것을 슬픔을 반영하는 행동으로서 더욱 유연하게 받아들이는 경향이 있기 때문이었다. 따라서 이러한 항목을 두 국가 간 연구에 그대로 사용하는 경우 국가간 우울에 대한 정도를 비교할 수 없게 된다. 또한 국내 연구에서도 해외의 측정항목을 국내로 들여와 사용함에 따라 생기는 측정상 문제로 오류가 발생할 수 있음을 인식하고, 이에 대한 연구가 필요함이 지적된 바 있었다(박원우, 김미숙, 정상명, & 허규만, 2007). 이처럼 측정동등성이 연구결과의 해석에 큰 영향을 미침에도 불구하고, 연구자들은 측정동등성에 대한 문제의식을 지니지 못한 채, 측정동등성을 가정하고 상이한 집단 간의 비교연구를 진행하는 경우가 많다. 측정 비동일성(measurement noninvariance)이 존재하는 경우 통계적 절차를 통하여 그 영향력을 통제하는 등의 조치를 취하는 것이 바람직하므로, 연구자는 반드시 사전에 이에 대해 검증할 필요가 있다.

따라서 본 연구는 이러한 중요성에도 불구하고 국내에서 그 방법론적 중요성이 충분히 인식되지 못하고 있는 측정동일성 문제를 다룸으로써 다음과 같이 경영학 분야의 발전에 기여하고자 한다. 첫째, 기존 국내 경영학 분야의 연구들이 아직 대부분 측정동등성의 이슈를 인식하지 못하고 있으며, 특히 방법론적으로 이 문제를 중점적으로 다룬 연구가 아직 존재하지 않는 상황에서, 본 연구는 최초로 국내 연구자들에게 측정동등성의 문제를 제기하고 활용을 안내하는 가이드의 역할을 제공할 수 있다. 둘째, 저자들은 측정동등성의 개념적 이해 뿐 아니라 그 활용실태에 대한 실증적 분석도 제공함으로써 국내 실상을 파악하고, 연구자들에게 경각심을 불러일으키는 역할을 할 수 있다. 셋째, 본 연구는 단순한 문제제기와 실태 파악을 넘어서서 구체적인 측정동등성 문제의 해결방법을 안내함으로써, 국내 연구자들의 향후 실증연구에 실질적인 공헌을 하고자 한다.

이에 본 연구는 우선 측정동등성에 대한 이해를 도모하기 위해 그 의미와 영향을 알아

보고, 측정동등성을 검증하지 않을 때 발생하는 문제점 및 국내 실태를 파악한 후, 측정동등성을 검증하는 여러 가지 방법을 제시한다.

Ⅱ. 측정동등성의 의미

1. 측정동등성의 의미와 필요성

앞서 언급한 바와 같이 측정동등성은 상이한 상황에서 동등한 속성을 측정할 때 고려해야하는 이슈이다. 여기서 '상이한 상황'이라 함은 시간(예: 근무시간), 인구적 요소(예: 문화, 평가자 집단), 혹은 측정매체(예: 웹 기반의 설문과 종이를 이용한 설문)의 차이를 말하며, 측정동등성은 이러한 상황적 상이함에 관계없이 측정의 안정성이 유지되는 것을 의미한다.

이러한 이질적 상황에 속해있는 복수 집단의 잠재 구성체(latent construct)에 대한 연구는 계속 증가하고 있으며, 연구자들은 비교연구를 위해 한 문화권에서 개발된 연구방법 또는 틀(framework)을 다른 문화권에 가져다 쓰거나(Bagozzi, 1994), 상이한 집단 간의 비교를 위해 하나의 척도를 두 집단에 모두 적용하게 된다. 이 경우 발생하는 집단 간 척도의 차이는, 서로 다른 집단에 속해있는 피평가자에 각각 내재한 측정된 잠재변수의 구조적 관계(structural relations)에 근본적 차이가 존재하기 때문이어야 한다. 그러나 이러한 측정 척도의 차이는 실질적으로는 다음과 같은 원인들에 기인할 수도 있다: (1) 서로 다른 집단에 속한 평가자들의 특정 측정도구를 해석하고 응답하는 방식의 차이, (2) 개별 평가자가 동일한 수행영역에서 동일한 대상을 평가할 때 수행을 동일하게 정의하지 않는 데에서 오는 체계적 편차, (3) 척도화 가공물(scaling artifacts), (4) 척도의 신뢰도 차이, (5) 연구된 구성체의 비동일성(nonequivalence of the constructs involved) (황호종, 2004; Steenkamp & Baumgartner, 1998; Vandenberg & Lance 2000). 따라서 이러한 오류를 방지하고 연구의 타당성을 확보하기 위해서는 잠재구성체 간 관계에 대한 가설검증 이전에 구성체를 측정하기 위한 측정도구(instrument)가 다른 문화, 혹은 집단에 적용가능한지에 대한 측정동등성을 검증하는 것이 필수적이다(황호종, 2004; Hui & Triandis, 1985; Steenkamp & Baumgartner, 1998).

그러나 일반적으로 실증연구의 수행에 있어 측정동등성에 대한 관심은 그 중요성에 비해 심각할 정도로 부족하다. 이에 대해 다음과 같은 원인이 제시되고 있다: (1) 측정동등성의 종류에 대한 합의 부족, (2) 다양한 측정동등성을 다루는 관련 용어의 일치성 미흡, (3) 측정동등성을 다루는 모형에 대한 이해 부족, (4) 측정동등성을 테스트하는 방법론의 복잡성, (5) 유의미한 집단 간 비교연구를 위해 필요한 측정동등성 범위에 대한 이해 부족, (6) 측정동등성을 증명하기 위한 방법에 대한 합의 부족 (Steenkamp & Baumgartner, 1998, p. 79).

2. 측정동등성이라는 용어의 고찰

측정동등성에 대한 국내외 연구에서 용어의 합의를 찾아보기 어려웠다. 먼저 국외에서는 영문으로 ME(measurement equivalence)나 MI(measurement invariance)로 주로 사용되었으나 학자에 따라서 이 둘을 혼용하여 쓰는 경우도 있으며, 비슷한 개념을 다른 형태로 표현하는 경우도 있었다(〈표 1〉 참조).

한편, 각종 문헌에서 측정동등성의 여러 측면에 대한 분류가 제시되고 있는데, 예를 들어 Mullen(1995)은 측정동등성을 번역동등성(translation equivalence), 측정간격동등성(metric equivalence), 환산동등성(calibration equivalence)으로 구분하여 이 세 가지를 통해 측정동등성을 만족시킬 수 있다고 하였다. 그 외에도 여러 문헌에서 측정동등성의 개념을 그 세부적인 동등성의 대상이나 사용 문맥에 따라 다음과 같은 여러 차원으로 제시된 바 있다: 어의동등성(semantic equivalence; 번역과정에서 단어나 문장의 선택이 원문에서 사용된 언어의 의미를 반영하는 정도), 내용동등성(content equivalence; 개념이나 문항이 그것이 사용된 문화적 환경과 관련하여 적절한지의 정도), 개념동등성(conceptual equivalence; 사용된 단어와 무관하게, 원문과 번역된 언어에서 특정 개념이 같은 형태로 존재하는 정도), 규범동등성(normative equivalence; 연구자가 사회적 규범의 차이에 의해 겪게 되는 문제의 정도. 어떤 문제가 사회적으로 논의될 수 있는지 터부시되는지 등이 있음), 기준동등성(criterion equivalence; 각 문화의 규범에 견주어서 주어진 변수의 측정결과가 동일하게 해석되는 정도), 기술적동등성(technical equivalence; 인터뷰나 지필고사 등 측정의 방법이 유사하게 인식되는 정도) (Flaherty, 1987; Cella et al., 1998; Singh, 1995; Cella, Lloyd & Wright,

1996; Touw-Otten & Meadows, 1996). 그러나 이와 관련하여 개념의 혼재가 아직 존재하고 있으며(Herdman, Fox-Rushby & Badia, 1997), 본 연구의 논의를 벗어나는 범위이므로 본 연구에서는 이러한 차원을 별도로 구분하지 않고 종합적으로 접근하여, 척도가 같은 개념을 측정하고 이집단 간의 근본적인 특질의 차이를 고려하였는지 여부를 측정동등성으로 사용하도록 한다(Smith, 2004).

〈표 1〉 국외에서의 측정동등성 용어의 사용 실태

용어	학자
ME	Drasgow(1984, 1987) Reise, Widaman, & Pugh(1993) Robie, Zickar, & Schmit(2001) Azevedo, Drost, & Mullen(2002) Raju, Laffitte, & Byrne(2002) Schneider, Hanges, Smith, & Salvaggio(2003) Liu, Borg, & Spector(2004) Mead, Lautenschalager, & Hecht(2005) Woehr, Sheehan, & Bennett(2005) Cole, Bedeian, & Field(2006)
MI	Meredith (1993) Reise, Widaman, & Pugh(1993) Schmit & Ryan(1993) Steenkamp & Baumgartner(1998) Bagozzi, Verbeke, & Gavino(2003) Reise & Henson(2003) Raykov(2004) Chen(2005) Reeve & Lam(2005)
ME/I	Vandenberg & Lance(2000) Vandenberg(2002) Meade & Lautenschlager(2004a, 2004b) Meade, Lautenschlager, & Hecht(2005)
혼용	Cheung & Rensvold(1999) Sharma & Weathers(2003)
기타	Lastovicka (1982)[2]

2) Lastovicka (1982)는 측정타당성의 7개 개념 중 하나로 invariance를 제시하였다.

국내에서는 측정동등성과 유사한 개념으로 검사의 동등화라는 개념이 이순묵(1992)에 의해 연구되었다. 이후 김학수(1997), 조용래와 김정호(2002)의 연구에서는 측정동일성이라는 용어를 사용하고 이를 검증하였으며, 황호종(2004)은 측정동등성이라는 용어의 사용과 함께 그 검증에 대한 방법론을 제시하였다.

본 연구에서는 측정동일성 혹은 측정동등성이 집단 간 개념에 대한 인식이 엄격하게 똑같지는 않더라도 통계적으로 허용할 만한 수준 내에서는 그 의미를 받아들일 수 있다는 의미에서 동일 및 동등의 사전적 의미[3]를 고려할 때 '동등'의 의미에 더 가깝다고 판단하였고, 따라서 '측정동등성'이라는 용어로 통일하여 사용하기로 하였다.

3. 측정동등성의 역사

집단 간 측정된 문항들이 여러 표본에 걸쳐 동일한지에 관한 문제는 오래전부터 Thomson과 Lederman(1939), Thurstone(1947) 등의 학자들에 의해 제기되어 왔으나, 이것이 요인동일성(factorial invariance)로 개념화되어 개별 문항이 문화 간에 동등하게 받아들여질 수 있는지, 만약 그렇지 않다면 어떤 문항이 동등하지 않은지에 대한 논의는 60년대부터 시작되었다(Meredith, 1964a, 1964b). 이후 측정동등성의 검증에 대한 연구가 시작되었는데, 1970년대 이전에는 다양한 휴리스틱 전략이 사용되었으나, 확인적 요인분석(Joreskog, 1971)과 문항반응이론(Lord, 1980)의 개발로 인해 휴리스틱 전략의 사용은 급격히 감소하게 되었다.

측정동등성에 관한 연구는 1990년대에 들어서면서 더욱 활발해졌는데, Meredith(1993)는 요인동일성(factorial invariance)과 더불어 측정동등성을 독립된 주제로 연구하였고, Labouvie와 Ruetsch(1995)는 측정 척도의 동등성 검증에 대한 연구를 실시하였다. 또한 국가 간 연구에 있어 측정동등성의 중요성이 활발하게 다루어졌다(Cella, Lloyd, & Wright, 1996; Holzmuller & Salzberger, 1999; Steenkamp & Baumgartner, 1998; Touw-Otten & Meadows, 1996)

이후 2000년대에 접어들면서는 지속적인 연구와 더불어(Robie, Zickar and Schmit, 2001; Azevedo, Drost, and Mullen, 2002) 기존 연구에 대한 이론적 검토 및 이후

3) 동일: 어떤 것과 비교하여 똑같음; 동등: 등급이나 정도가 같음. 또는 그런 등급이나 정도.

연구에 대한 제언 뿐 아니라 그 검증방법과 해결책에 대한 통합적 리뷰가 이루어지는 등 연구에 대한 발전적 논의가 이루어졌다(Vandenberg, 2002; Vandenberg & Lance, 2000; Raju, Laffitt, & Byrne, 2002; Spini 2003).

4. 유사개념

측정동등성의 유사개념으로는 검사의 동등화(test equating)와 에틱타당성(etic validity)이 있다. 검사의 동등화는 "상이한 두 양식 간에 동등한 의미를 가지는 점수들끼리의 짝을 찾아내기 위한 통계적 절차를 의미한다"(이순묵, 1992, p. 165). 같은 속성을 측정하기위해 동시에 두 개 이상의 양식을 개발하거나, 같은 검사 프로그램에서 매년 새로운 양식을 개발하는 경우와 같이, 같은 검사에 대해 다수의 양식이 개발될 경우, 하나의 원점수가 모든 양식에 정확히 같은 의미를 가지는 것을 보장할 수 없기 때문에, 공정한 검사를 위해서는 상이한 양식들이 동등화되어야 한다. 검사의 동등화를 위한 통계적 절차로서 동일백분위방식, 선형적 동등화, 확인적 요인분석을 사용할 수 있다(이순묵, 1992).

에틱타당성은 경영학의 분야들 중 이문화(異文化) 마케팅 연구(cross-cultural/national marketing research)나 국제경영 등의 분야에서 중요한 방법론적 이슈로 다루어지는 개념으로, 인류학의 에믹-에틱 딜레마(emic-etic dilemma)라는 개념에서 유래된 것이다. 에믹은 특수한 언어나 문화 내에서만 의미가 있는 내부자적 관점이다. 반면, 에틱은 관찰자 혹은 외부자적 관점에서 다른 사회 문화 체계를 기술하고 비교하는 분석적 관점, 혹은 검증할 수 있는 과학적 판단을 말한다. 에틱 모형은 비교와 일반화를 수립하기 위해서 각 문화들 간의 경계를 초월하여 보편성의 수준에서 접근하는 것인 반면, 에믹 모형은 에틱 모형의 사용과 같은 과학적 일반화는 불가능하고 특정 문화의 서술과 해석의 작업만이 가능하다는 접근방식이다(Berry, 1980). 따라서 에틱타당성은 여러 집단에서의 비교 및 일반화 가능성을 제시하는 외적타당성(external validity)을 의미한다.

Ⅲ. 측정동등성 미검증시의 문제점

1. 문제점

국가, 문화 그리고 집단 간 비교연구에서 연구자들은 측정 개념에 대해 비교, 일반화할 수 있는 동등한 속성을 가정한다. 만일, 비교하고자 하는 속성이 상이한 집단 간에는 비교될 수 없는 것이거나 비동일한 것이라면 연구의 결과는 타당성을 잃게 된다. 그러나 대부분의 연구에서는 동등한 속성을 가정할 뿐, 그것이 정말 동등한 속성인지에 대한 검증을 하지 않는 측정동등성의 문제를 보인다.

앞서 언급한 에틱-에믹 패러다임에서 보면, 에틱타당성이 확보되기 전까지는 에믹 모형을 가정하는 것이 바람직하나 대부분의 연구에서 에틱타당성에 대한 검증 없이 자의적으로 에틱타당성을 가정하는 것이 측정동등성의 문제에 해당한다. 이를 "부과된 에틱타당성(imposed etic validity)"이라고 하고, 이 부과된 에틱타당성이 실제로 입증될 경우에 이를 "도출된 에틱타당성(derived etic validity)"이라고 한다(Berry, 1980). 도출된 에틱타당성이 확보될 때까지, 측정은 에믹 모형을 가정하는 것이 타당하다.

측정동등성이 확보되지 않을 경우, 에틱타당성과 같은 외적타당성 문제 외에도 예측타당성의 이슈가 발생한다(Chan & Rossiter, 2003; Voronov & Singer, 2002). 연구결과가 진정 연구자가 가정한 변수들 간의 관계에서 기인하는 것인지를 알 수 없기 때문에 연구의 예측타당성이 약화될 수밖에 없는 것이다. Steenkamp와 Baumgartner (1998)는 그들의 연구에서 Gaski와 Etzel(1986)의 연구결과를 재분석하여 측정동등성을 검증하지 않을 시 발생하는 예측타당성 문제를 지적하고 있다. Gaski와 Etzel(1986)의 유럽의 4개국의 소비자 간 광고에 대한 태도의 차이에 대한 분석 결과가 ANOVA의 원점수(raw score)를 비교했을 때는 $p<.05$ 수준에서 유의미하지 않았다. 하지만 이들이 제시하는 측정동등성 검증 절차에 따라 잠재평균값을 비교해본 결과 $p<.001$ 수준에서 그 차이가 유의미하게 나타났다. 이들의 분석 결과는 집단 간 비교연구에 있어서 측정동등성을 검증하지 않을 경우 그 연구의 결과가 달라질 수 있음을 보여주는 좋은 예이다. 따라서 변수들 간의 인과관계에 대한 이해, 설명 그리고 예측을 증진시키기 위해서는 측정동등성이 반드시 확보되어야 한다. 마지막으로, 측정동등성이 확보되지 않을 경우 연구의 결론타당성이 위협받게 된다는 문제점이 있다(Albaum & Baker, 2005;

Douglas & Craig, 1983; Green & White, 1976; Mullen, 1995).

2. 국내외 학계의 측정동등성 인식 정도

1) 국외 문헌에서의 측정동등성 인식 여부

국외의 방법론 책을 조사한 결과, 다변량 분석법이나 측정이론과 관련된 책들에서 측정동등성에 대한 언급 및 검증 방법의 소개, 유사개념이나 유사용어를 설명한 것을 확인할 수 있었다. 이를테면 Hair, Black, Babin, Anderson 그리고 Tatham (2006)의 *Multivariate Data Analysis*에서는 공분산구조모형과 확인적 요인분석 부분에서 측정이론과 관련하여, 다중표본 요인분석(multiple sample factor analysis)을 이용하여 측정동등성을 검증하는 단계에 대해 자세히 설명하고 있다. 또한 Shultz와 Whitney(2005)의 *Measurement Theory in Action*에서는 다양성 이슈(diversity issues)[4] 부분에서 측정동등성과 유사개념인 검사동일성(test equivalence)에 대해 설명하며, 문항반응이론 부분에서는 측정동등성의 반대개념인 측정편파(measurement bias, 검사 문항이 서로 다른 집단 하에서 측정하고자 하는 개념을 제대로 나타내지 못하는 것)에 대해 간단히 설명하고 있다.

국외 학술지의 경우 Academy of Management Journal과 Journal of Applied Psychology에 게재된 지난 5년간의 논문을 대상으로 측정동등성이 검증되어야 할 논문을 선정한 결과 각각 8개과 30개의 논문을 찾을 수 있었다. 분석 결과 AMJ에서는 8개의 논문 중 측정동등성을 인식하고 그 용어도 사용한 논문은 1개였으나 용어는 사용하지 않았더라도 인식은 하고 있는 논문이 2개 발견되었다. JAP의 경우 측정동등성을 인식하고 그 용어도 사용한 논문이 7개, 용어 사용여부와 무관하게 측정동등성을 인식하고 있는 논문은 12개로, 약 40%에 달하는 논문들이 측정동등성을 인식하고 있음을 알 수 있었다.

4) 민족성, 성, 나이, 언어 능력, 신체적 장애 등에 의해 구분지어지는 표본의 검사 시 발생하는 문제를 의미한다.

〈표 2〉 국외 학술지 논문의 측정동등성 인식 정도

학술지	저자 (연도)	비교내용	신뢰도	타당도	측정동등성	
					인식	용어사용
AMJ	Lam, Chen, & Schaubroeck (2002)	국가 간 연구	O	O	O	X
	Lester & Meglino (2002)	종단연구	O	O	O	X
	Chatman & O'Reilly (2004)	집단 간 연구	O	X	X	X
	Cullen & Parboteeah (2004)	국가 간 연구	O	O	X	X
	Spicer, Dunfee, & Bailey (2004)	국가 간 연구	O	O	X	X
	Perrewe, Zellars, Ferris, Rossi, Kacmar, & Ralston (2004)	집단 간 연구	O	O	X	X
	Chen (2005)	종단연구	O	O	O	O
	Polzer, Crisp, Harvenpaa, & Kim (2006)	지역적 차이	O	O	X	X
JAP	Gelfand, Higgins, Nishii, Raver, Dominguezm, Murakami, Yamaguchi, & Toyama (2002)	국가 간 비교	O	O	X	X
	Saks & Ashforth (2002)	종단연구	X	X	X	X
	Truxillo, Bauer, Campion, & Paronto (2002)	종단연구	O	O	X	X
	Cable & DeRue (2002)	잡단 간 연구	O	O	O	O
	Fuller, Stanton, Fisher, Spitzmuller, Russell, & Smith (2003).	종단연구	O	O	X	X
	Ambrose & Cropanzano (2003)	종단연구	O	O	X	X
	Simmering, Colquitt, Noe, & Porter (2003)	종단연구	O	O	X	X
	Bagozzi, Verbeke, & Gavino (2003)	국가 간 연구	O	O	O	O
	Donovan & Williams (2003)	종단연구	O	X	X	X
	Avery (2003)	집단 간 연구	O	X	X	X
	Shaw, Duffy, Mitra, Lockhart, & Bowler (2003)	종단연구	O	X	X	X
	Schneider, Hanges, Smith, & Salvaggio (2003)	종단연구	O	X	O	O
	Smither & Walker (2004)	종단연구	O	O	X	X

〈표 2〉 국외 학술지 논문의 측정동등성 인식 정도 (계속)

학술지	저자 (연도)	비교내용	신뢰도	타당도	측정동등성	
					인식	용어사용
JAP						
	Simpson & Stroh (2004)	집단 간 연구	O	X	X	X
	Liu, Borg, & Spector (2004)	종단연구	O	O	O	O
	Fullagar, Gallagher, Clar, & Carroll (2004)	종단연구	O	O	X	X
	Woehr, Sheeha, & Bennett (2005)	집단 간 연구	O	O	O	O
	Epitropaki & Martin (2005)	집단 간 연구 종단연구	O	O	O	X
	Westaby & Lowe (2005)	종단연구	O	O	X	X
	Begley & Lee (2005)	종단연구	O	O	X	X
	Bauer, Erdogan, Lide, & Wayne (2006)	종단연구	O	X	O	X
	Eddleston, Veiga, & Powell (2006)	집단 간 연구	X	O	O	X
	Gong & Fan (2006)	종단연구	O	O	X	X
	Johnson,, Morgeson, Ilgen, Meye, & Lloyd (2006)	집단 간 연구	O	X	X	X
	Tay, Ang, & Van Dyne (2006)	종단연구	O	O	O	O
	Fritz & Sonnetag (2006)	종단연구	O	O	O	O
	Porath & Bateman (2006),	종단연구	O	X	X	X
	de Jonge & Dormann (2006)	종단연구	O	O	X	X
	Keller (2006)	종단연구	O	O	O	X
	Shaffer, Harrison, Gregersen, Black, & Ferzandi (2006)	국가 간 연구 종단연구	O	O	O	X

2) 국내 문헌에서의 측정동등성 인식 여부

국내에서의 인식 수준을 알기위해 국내에 출판된 방법론관련 책(21권)을 조사한 결과, 측정동등성에 대한 언급이 전혀 없었으며, 관련 내용도 전무한 것으로 드러나 국내 방법론 학계의 경우 측정동등성에 대한 인식이 전혀 되지 않고 있는 것으로 밝혀졌다.

한편 학술지의 경우, 국내 주요 학술지인 *경영학연구*, *인사·조직연구*, *인사관리연구*에

게재된 지난 10년간의 논문을 대상으로 하여, 측정동등성에 대한 검증이 필요하다고 판단되는 논문을 선정하여 분석하였다. 그 결과, 총 17개의 논문을 찾을 수 있었는데, *경영학연구*에서는 10개 논문 중 측정동등성을 인식하고 있으며 그 용어를 사용한 논문은 1개뿐이었고, 인식은 하고 있으나 용어는 사용하지 않은 논문도 2개에 불과했다. *인사·조직연구*와 *인사관리연구*의 경우 그 인식의 정도가 더욱 미미하여 선정된 7개의 논문 중 측정동등성이라는 용어를 사용한 논문은 한 개도 없었으며, 용어를 사용하지 않았더라도 측정동등성을 인식하고 있는 논문도 단 한 개에 불과했다.

이렇듯 국내외 주요 학술지 및 방법론 서적의 측정동등성의 인식 수준을 분석해본 결과, 상당 수준으로 측정동등성에 대한 인식이 이루어져있고 이에 대한 논의가 이루어지

〈표 3〉 국내 학술지 논문의 측정동등성 인식 정도

학술지	저자 (연도)	비교내용	신뢰도	타당도	측정동등성	
					인식	용어사용
경영학연구	김학수 (1997)	집단 간 비교	O	X	O	O
	강정애 (1997)	조직 간 비교	O	X	X	X
	김경수 (1998)	집단 간 비교	O	O	X	X
	김규남, 신만수 (2001)	국가 간 비교	O	O	X	X
	이덕로, 서도원 (1998)	종단연구	O	O	X	X
	김정구, 김태웅, 박승배 (2003)	집단 간 비교	X	X	X	X
	서문식, 김상희 (2004)	집단 간 비교 종단연구	O	O	X	X
	장은주, 박경규 (2005)	집단 간 비교	O	O	X	X
	엄명용, 김태웅 (2006)	집단 간 비교	O	O	O	X
	박철, 이태민 (2006)	국가 간 비교	O	O	X	X
인사·조직 연구	강혜련 (1998)	집단 간 비교	O	O	X	X
	김영조 (2000)	종단연구	O	O	X	X
	이지우, 김종우 (2002)	집단 간 비교	O	O	X	X
	권순식, 김상진 (2005)	집단 간 비교	O	O	X	X
인사관리연구	장동운 (2003)	집단 간 비교 국가 간 비교	O	O	X	X
	김정원, 김태형, 권중생 (2004)	집단 간 비교	O	O	X	X
	이태식 (2005)	국가 간 비교	O	O	O	X

고 있는 해외와 비교하여, 국내 경영학계는 측정동등성에 대한 연구가 일천하고 연구자 간에서도 이에 대한 인식의 수준이 미미함이 드러났다. 이에 다음과 같은 시사점을 도출할 수 있었다. 우선 국내 경영학 분야 내에 측정동일성에 대한 인식을 확산시킬 필요성이 있음을 알 수 있다. 비록 극히 일부의 연구가 측정동등성을 인식하고 연구 내에서 이 용어를 직접적으로 사용하고 있기는 하나, 아직은 그 정도가 크게 미흡한 것이 본 분석을 통해 밝혀졌기 때문이다. 또한 측정동등성의 개념을 알리는데 그치지 않고, 이를 국내 연구자들이 적극적으로 연구에 반영하여 연구의 타당성을 높일 수 있도록 극복을 위한 구체적인 해결책을 제시할 필요가 있음을 알 수 있었다.

Ⅳ. 측정동등성의 검증방법

앞서 논한 바와 같이 측정동등성은 많은 연구에서 필수적으로 검증되어야 하는 전제조건임에도 불구하고 국내 학계에서는 그 인식정도가 크게 미흡한 현실이다. 따라서 본 장(章)에서는 측정동등성 검증을 위해 사용되는 대표적 방법과 그 적용방법을 간략히 제시하여 국내 연구자들이 측정동등성을 검증하고 이를 통해 타당성 있는 연구를 하는데 도움이 되고자한다. 이를 위해 여기서는 측정동등성 검증의 가장 대표적 방법으로, 상호보완적으로 병행되어 쓰이는 두가지 방법인 확인적 요인분석과 문항반응이론에 대해 다루고(Raju et al, 2002), 보조적으로 일반화가능도이론에 대해 설명하도록 하겠다.

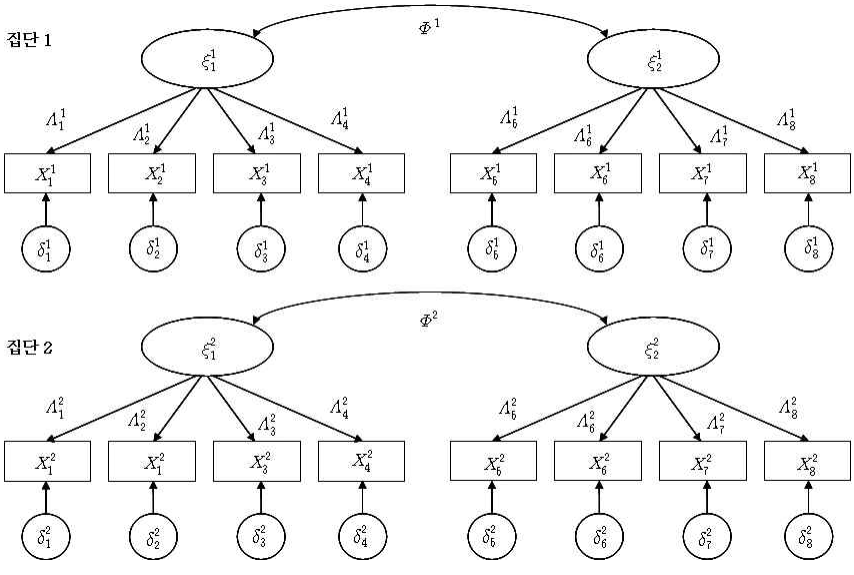
1. 확인적 요인분석 (Confirmatory Factor Analysis, CFA)

1) 개념과 연구방법

CFA는 "측정된 변수들이 구성체(construct)를 얼마나 잘 나타내는가를 검증하는 방법"이다(Hair et al., 2006, p. 773). CFA는 연구자의 이론적 배경을 바탕으로 한 모형이 얼마나 실제 자료와 일치하는 지를 분석하여 통계치로 나타내며, 이를 통해 연구자의 연구 모형을 확증하거나 기각할 수 있게 해준다(Hair et al., 2006). 따라서 CFA를 활용하면 연구자가 설계한 측정모형이 지지될 수 있는지, 혹은 수정이 필요한지를 판단할 수 있다.

2) CFA와 측정동등성: 개념적 접근

관찰된 변수와 가설적 구성체 사이의 관계에 대한 연구자의 측정이론은 CFA 모형을 사용해 나타낼 수 있다. 〈그림 1〉은 두 집단에서의 두 개의 구성체 측정모형에 대한 CFA 검증을 도식화한 것이며 이와 같은 측정모형에서 연구자들은 일반적으로 세 가지 가정을 한다. 먼저, 각 집단 간 잠재변수(latent ξ)들의 밑바탕이 되는 개념의 동질성을 가정한다. 둘째로, 집단 간 측정변수(X)와 잠재변수(ξ) 간의 동등한 관계(λ)를 가정하며, 마지막으로 집단 간 측정변수(X)는 집단 간 동일한 유일인자(unique factor, δ)에 의해 동일한 정도의 영향을 받는다고 가정한다(Bollen, 1989; Drasgow, 1984, 1987; Vandenberg & Self, 1993). 연구자들은 실제 비교연구에서 위의 세 가지 가정들에 대한 검증없이 MANOVA나 ANOVA를 이용하여 집단 간 평균 차이를 검증하는데, 이 경우 집단 간 관찰된 차이가 집단의 차이로 인한 것인지, 혹은 측정의 비동일성으로 인한 것인지 확인할 수 없다. 따라서 집단 간 평균 차이를 검증하기 전에 CFA를 이용한 측정동등성을 확인할 필요가 있다.



〈그림 1〉 CFA 모형

〈그림 1〉에서 나타낸 측정동등성 문제를 g번째 집단의 k 문항들에 대한 관계를 식으

로 나타내면 다음과 같다(Vandenberg & Lance, 2000).

$$X_k^g = \tau_k^g + \Lambda_k^g \xi^g + \delta_k^g \quad (1)$$

X_k^g : 측정변수
τ_k^g : 회귀식 절편의 벡터
Λ_k^g : 회귀식에서의 기울기
ξ^g : 잠재변수
δ_k^g : 유일인자의 벡터

만약 ξ^g와 δ_k^g의 기대값이 0이라고 가정한다면 공분산 식은 다음과 같다.

$$\Sigma^g = \Lambda_k^g \Phi^g \Lambda_k^{g'} + \Theta_\delta^g \quad (2)$$

Σ^g : k item 중 g번째 집단의 분산과 공분산행렬
Λ_k^g : 요인 부하량행렬
Φ^g : 잠재변수들 간 분산과 공분산
Θ_δ^g : 오차분산(unique variance)의 대각행렬

CFA를 바탕으로 한 위의 식 (1)과 (2)를 통해, 집단 간 잠재변수, 회귀식의 기울기, 회귀 절편, 요인 부하량 행렬, 오차분산, 그리고 잠재변수들 간 분산과 공분산의 동일성 여부를 검증할 수 있다.

3) CFA와 측정동등성: 실증적 접근

앞 절에서의 〈그림 1〉과 식(1), (2)를 바탕으로 다음과 같은 구체적인 검증방법을 도출해 낼 수 있다.

(1) 검증방법 1: 공분산행렬동일성(equivalent covariance matrices 혹은 omnibus test) 검증

집단 간 공분산행렬이 동일하다($\Phi^g = \Phi^{g'}$)는 영가설을 검증한다. 그러나 공분산행렬의

동일성 검증에 대해서는 합의가 없으며(Raju et al., 2002) 이 검증의 유용성, 진단가치에 의문이 제기되어 왔다(Meade & Lautenschlager, 2004a). 따라서 연구자들은 일반적으로 이 검증의 결과와 상관없이 다음 검증을 진행한다(Hair et al., 2006).

(2) 검증방법 2: 형태동일성(configural invariance 혹은 factor structure equivalence) 검증

형태동일성 검증은 공분산행렬 검증에 의해 동일성이 검증되지 않을 경우 동일성 결여의 원인을 결정하기 위해 실행된다. 형태동일성 검증에서는 집단 간 요인구조만이 제약을 받으며, 각 표본에서 λ, Φ, θ가 자유롭게 추정되기 때문에 이 값들이 집단 간 차이가 있을 수 있다. 이 때 χ^2통계치는 요인구조를 제외한 다른 제약을 받지 않으므로 $\chi^2_{무제약}$로 나타낼 수 있으며, 기저모형(baseline model)의 역할을 한다. χ^2값과 적합도는 연구자의 모형이 집단들의 공분산행렬에 얼마나 적합한지를 나타내며, 만약 집단 간 CFA의 적합도 지수가 적절하다면, 최소한의 크로스확인(cross validation)에 대한 증거를 확보한 것이라고 볼 수 있다(Hair et al., 2006).

(3) 검증방법 3: 측정간격동일성(metric invariance) 검증

집단 간 요인 부하량이 동일하다($\Lambda_k^g = \Lambda_k^{g'}$)는 영가설을 검증한다. 연구자의 측정이론이 서로 다른 집단의 응답자에 의해 같은 방식으로 사용된다는 것으로, 이는 집단 간 응답자가 평정 척도(rating scales)를 유사하게 사용한다는 것을 의미한다(Hair et al, 2006). 이 완전제약 하의 모형에서 통계치는 $\Delta\chi^2$(즉, $\chi^2_{제약} - \chi^2_{무제약}$)와 적합도를 사용하는데, $\Delta\chi^2$값이 유의미하다면 기저모형이 자료에 더 적합하다는 것으로 측정의 동일성이 존재하지 않는다는 것을 의미하게 된다. 측정간격동일성은 이후 설명할 절편동일성과 잠재변수평균동일성의 전제조건이 된다(Cheung & Rensvold, 1999).

(4) 검증방법 4: 절편동일성(scalar invariance) 검증

잠재변수에 대한 문항의 회귀식 절편이 집단 간 동일하다($\tau_k^g = \tau_k^{g'}$)는 영가설을 검증한다. 형태동일성이나 측정간격동일성은 공분산에 대한 정보만을 요구하나, 많은 연구에서는 집단 간 평균을 비교하는 것 또한 매우 중요한데, 이를 위해서는 절편동일성을 검증

해야 한다(Meredith, 1993). 절편동일성이 검증되면 측정된 문항의 집단 간 평균차이가 잠재적 변수의 평균의 차이로 인한 것임을 알 수 있다(Steenkamp & Baumgatner, 1998). 절편동일성은 측정간격동일성과 마찬가지로 $\Delta\chi^2$와 적합도에 의해 확인될 수 있다.

(5) 검증방법 5: 오차분산동일성(invariant uniquenesses) 검증

집단 간 동일 문항의 오차분산이 동일하다($\Theta_\delta^g = \Theta_\delta^{g'}$)는 영가설을 검증한다. 이 검증은 연구자가 집단 간 척도(scale)의 신뢰도를 비교하고자 할 때 실행한다(Cheung & Rensvold, 1999).

(6) 검증방법 6: 요인분산/공분산동일성(invariant factor variances / covariances) 검증

요인분산과 공분산동일성은 집단 간 요인의 분산이 동일($\Phi^g = \Phi^{g'}$)하다는 영가설과 공분산이 동일($\Phi_j^g = \Phi_j^{g'}$)하다는 영가설을 검증하는 것이다. 요인분산/공분산동일성은 연구자가 집단 간 잠재변수들의 상관관계를 비교하고자 할 때 실행된다(Byrne, 1994; Jackson, Wall, Martin, & Davids, 1993; Marsh, 1993)

(7) 검증방법 7: 잠재변수평균동일성(equal factor means) 검증

집단 간 요인평균이 같다는 ($\Phi_{jj}^g = \Phi_{jj}^{g'}$) 영가설을 검증한다.

CFA 분석을 통한 검증에서는 측정동등성을 좀 더 포괄적으로 약한 요인동일성(weak factorial invariance), 강한 요인동일성(strong factorial invariance), 엄격한 요인동일성(strict factorial invariance)로 구분하기도 한다(Meredith, 1993). 약한 요인동일성은 집단 간 요인 부하량값이 불변한다는 것만을 검증하며, 강한 요인동일성에서는 집단 간 요인 부하량 뿐만 아니라 측정된 변수의 절편도 불변한다는 것을 검증한다. 마지막으로 엄격한 요인동일성에서는 집단 간 요인부하량, 요인의 절편, 오차분산이 모두 불변하다는 검증으로, 잠재변수들의 관계를 제외한 모든 값이 같아야 함을 의미한다.

Cheung과 Rensvold(1999)는 측정동등성을 검증했을 때, 동등하지 않은 문항들이 관찰된다면 다음의 세 가지 방법을 통해 해결할 수 있다고 한다. 첫번째로 동일성이 없

는 문항을 제거하는 것이다. 모형으로부터 문항을 제거하는 것은 특정 연구에 대해 유용성을 증가시킬 수 있겠지만, 다른 이론을 설명하기에 적합하지 않으므로 자료의 수만큼 모형이 증가하게 되며 이는 절약의 원리(principle of parsimony)와는 맞지 않는다는 한계가 있다.

다음으로 부분적 요인동일성(partial factorial invariance, PFI) 검증을 사용하는 것이다. 부분적 요인동일성은 동일한 문항의 부하량은 집단 간 동일하다는 제약을 받지만, 동일하지 않은 문항의 부하량은 서로 차이가 나도록 둔다. Byrne, Shavelson과 Muthen(1989)은 측정동등성을 실질적으로 적용 시 완전한 동일성을 유지하는 것이 쉽지 않은 경우가 종종 있기 때문에 완전동일성이 성립하지 않을 경우, 연구자들은 적어도 부분적 측정동등성이 존재하는지에 대해서는 확신해야 한다고 했다(Steenkamp & Baumgatner, 1998). 이 때 동일하지 않은 문항이 모형의 매우 작은 부분을 차지하며 따라서 집단 간 비교분석에 있어서 중요할 만큼의 영향력이 없다는 것을 가정한다. 연구자들은 동등하지 않은 문항을 확인하거나 동일성에서 벗어난 정도를 확인하고 부분 동일성의 적절성을 밝혀야 한다(Cheung & Rensvold, 1999).

마지막으로, 검증된 비동일성을 집단 간 유의미한 차이를 나타내는 정보의 원천으로 취급할 수도 있다. 만약 어떤 측정에 있어서 번역 상 언어적 차이로 인해 비동일성이 나타난 것이 아니라 문화, 국가적 차이로 인한 것이라면 연구자는 응답간 비동일성을 실제 문화 간 차이로 여기고 조사할 수 있다(Cheung & Rensvold, 1999).

2. 문항반응이론 (Item Response Theory, IRT)

1) IRT의 개념과 연구방법

기존의 고전적인 문항분석통계(item analyses statistics)는 어떠한 특성(trait)에 있어 다른 수준의 능력을 가진 피응답자들이 각 문항들에 대해 어떻게 반응하는지에 대한 정보를(즉, 문항함수를) 제공하지 못한다는 한계를 지니고 있다. 이에 반해 IRT는 문항함수(item functions)가 어떻게 이루어져있는가에 대한 정보를 제시한다(Crocker & Algina, 1986). 즉, 이는 고전검사이론(traditional test theory)처럼 검사가 단순히 문항들의 총점에 의해 분석되는 것이 아니라, 문항 하나하나의 독특한 특성을 지닌 고유한 문항특성곡선(item characteristics curve)에 의해 분석된다는 이론이다(성태제, 2001). 특히 이는 모

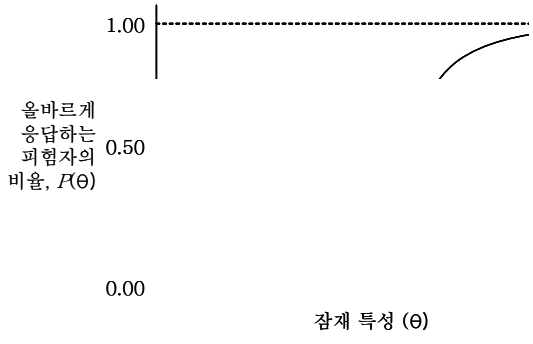
든 문항들이 동일한 능력을 측정한다는 단일차원성(unidimensionality)과 국지적 독립성(local independence)이라는 두 가지 가정을 기반으로 하며, 문항의 특성을 나타내는 문항모수들과 응답범주 각각을 선택할 수 있는 가능성에 대한 개인들의 특성을 나타내는 잠재특성들과 연관되어 있다. 또한 이들 간의 확률적 관계는 수학적으로 비선형적인 문항응답함수(item response function, IRF)로 나타내진다(Chernyshenko, Stark, Chan, Drasgow, & Willianms, 2001). 이러한 IRT는 본래 이분(dichotomous)점수의 문항들에 대한 이해를 향상시키기 위해 개발되었다. 하지만 후에 등급반응모형(graded response model: Samejima, 1969)을 이용하여 리커트 척도(Likert scale)와 같은 다분(polytomous) 문항들에서도 적용이 가능하게 되었다 (Meade & Lautenschlager, 2004b).

2) IRT와 측정동등성: 개념적 접근

(1) 문항특성곡선(Item Characteristics Curve, ICC)

IRT를 이해하기 위해서는 문항특성곡선에 대한 이해가 필수적이다. 문항특성곡선이란 "문항에 대해 올바르게 응답하는 확률을 검사 문항들의 성과를 나타내는 잠재특성(latent trait, Θ)의 함수로 나타낸 것"이다(Crocker & Algina, 1986, p.340). 즉, 이는 IRT 모형에서 다양한 수준의 능력을 측정하기 위해 주어진 문항들에 대해 응답자들이 정확하게 답하는 확률을 나타내기 위한 수학적 함수로(Shultz & Whitney, 2005), 준거변수(criterion variable)와 한 문항에 답을 맞힐 확률과의 함수적 관계를 말한다(성태제, 2001). 검사개발에 많이 쓰이는 S모양의 문항특성곡선은 다음의 〈그림 2〉와 같다.

IRT에서 각 문항은 〈그림 2〉와 같은 그래프를 도식화 할 수 있도록 하는 모수들의 집합으로 표현된다. 모수들의 수에 따라 각기 1, 2, 3 모수모형으로 나뉘는데 본 연구에서는 가장 널리 쓰이고 있는 2 모수모형을 기준으로 설명하도록 하겠다. 먼저 모수 a는 잠재특성 θ와 주어진 잠재특성의 수준에서 응답의 조건적인 확률 $P(\theta)$사이의 관계의 기울기에 대한 비율을 의미한다. 이 모수는 문항이 각 개인들과 다른 수준의 θ사이를 판별하는 정도를 반영한다. 다음으로 모수 b는 잠재특성과 응답가능성 사이의 관계를 모형화하는 역할을 한다. 즉, 모수 a는 문항특성곡선의 잠재특성과 관찰점수 사이의 관계를 나타내는 기울기를(즉, 문항변별도를), 모수 b는 주어진 문항특성곡선의 굴곡점(inflexion point)



자료원: Crocker & Algina, 1986, p341

〈그림 2〉 문항특성곡선의 예

에 기반한 문항특성곡선의 수평적 위치를 결정하는 역할을 한다. IRT 모형에서 모수 b는 문항난이도 모수(item difficulty parameter)라고 불린다(Meade & Lautenschlager, 2004b).

(2) DIF(Different Invariance Functioning)

문항편이(item bias)와 측정동등성은 Lord(1980)에 의해 처음으로 제시되었다. 그는 만약 문항들이 모든 집단에서 같은 문항특성곡선을 따른다면 설문응답자들에 대해 측정동등성이 유지된다고 주장했다. 이 조건은 이분문항응답을 위한 IRT에서 검사가 측정하는 특성 θ에 대해 다른 집단의 구성원들이 주어진 응답을 할 가능성 $P(\theta)$가 동등하다는 것을 의미한다(Mellenbergh, 1989; Millsap & Everson, 1993). 즉, 참가자 i에 의해 응답된 문항 j, 그리고 잠재특성 T를 측정한다고 가정할 때, 그리고 변수 v를 선택하는 것과 관련된 측정동등성은 문항응답 x_{ij}의 분포함수 F와 동일할 때 다음과 같다(Borsboom, Mellenbergh & Heerden, 2002, p.434). 모든 $\{x, t, v\}$에 대해,

$$F(X_{ij}=x_{ij} \mid T=t_i, V=v_i) = F(X_{ij}=x_{ij} \mid T=t_i) \tag{3}$$

어떠한 상황에서건 위의 조건에 벗어나면, 즉, 다음 아래와 같은 조건이 되면 편이된(biased) 문항이라고 말할 수 있다.

$$F(X_{ij}=x_{ij}|\ T=t_i, V=v_i) \neq F(X_{ij}=x_{ij}|\ T=t_i) \quad (4)$$

이러한 상황을 IRT문헌에서는 DIF(different invariance functioning)라 부른다. 이는 절대측정(absolute measurement)과 상대측정(relative measurement)의 개념을 이용하여 좀 더 자세히 설명될 수 있다. 절대측정이란 절대척도(absolute scale) 상의 특성을 측정하는 과정을 의미한다(예: "나는 도서관에서 책장 위쪽에 있는 책을 꺼내는데 곤란을 겪는다."). 즉, 이는 문항응답자가 속한 집단이나 상황 등의 상대적인 위치와 상관없이 동등하게 인식되는 측정으로 동일한 측정이 연구대상의 모든 구성원들에게 동일하게 적용하게 된다. 반면 상대측정은 상대척도(relative scale) 상의 특성을 측정하는 과정을 의미하는데(예: "나는 농구 팀에서 잘 할 수 있다.")(Borsboom et al., 2002, p. 435), 이는 응답자가 속해있는 집단의 상대적인 위치에 따라 상이한 측정이 이루어지는 것이다. 이러한 다른 형태의 측정은 각기 다른 측정동등성과 편이를 암시한다.

〈표 3〉 절대측정과 상대측정의 측정동등성 유지와 편이의 비교

	절대측정	상대측정
측정동등성 유지	$F(X_{ij}=x_{ij}\mid T=t_i, V=v_i) = F(X_{ij}=x_{ij}\mid T=t_i)$	$F(X_{ij}=x_{ij}\mid W=w_i, V=v_i) = F(X_{ij}=x_{ij}\mid W=w_i)$
편이	$F(X_{ij}=x_{ij}\mid T=t_i, V=v_i) \neq F(X_{ij}=x_{ij}\mid T=t_i)$	$F(X_{ij}=x_{ij}\mid W=w_i, V=v_i) \neq F(X_{ij}=x_{ij}\mid W=w_i)$

자료원: Borsboom, Mellenbergh & Heerden, 2002, pp. 435-436 재구성

여기서 W는 특성의 상대적인 위치를 의미하는데, 이는 문항에 응답하는데 포함되는 응답자의 인지적 과정에 따라 달라진다. 따라서 Borsboom 등(2002)은 W는 대부분의 검사에서 알려지지 않지만 특성 T의 변환으로 이루어져야 하며, 문항이나 검사에 따라 달라질 수 있다고 했다. 또한 그들은 위의 식을 통해 절대측정동등성과 상대측정동등성은 오직 이들의 모집단 분포의 평균과 분산에 차이가 없을 때만이 동시에 존재할 수 있다고 밝혔다. 만약 분포의 평균이나 분산에 차이가 있을 경우 절대측정은 상대편이(relative bias)를 일으키게 되고, 상대측정동등성은 절대편이(absolute bias)를 일으킬 것이기 때문이다. 즉, 이는 일반적인 척도에 대한 개인간 혹은 집단 간 평균 차이의 타당한 비교를 위해서는 문항들이 측정동등성을 수반해야 한다는 것을 의미한다. 다른

말로 특정 하부집단들에 대한 편이가 존재하지 않아야 한다(Reise & Henson, 2003).

3) IRT와 측정동등성: 실증적 접근

(1) IRT 기반의 DIF 검증방법

IRT 기반의 측정동등성의 검증은 사용되는 특정한 방법론의 형태에 따라 달라진다. IRT를 기반으로 하는 DIF검증방법은 여러 가지가 존재하는데, 먼저 Lord(1980)의 chi-square χ_j^2, Raju(1988, 1990)의 two area measures $Z_j(ESA)$, $Z_j(H)$, Thissen, Steinberg와 Wainer(1988)의 LR검사(likelihood ratio test), Raju, van der Linden & Fleer(1995)의 DFIT(differential functioning of items and tests) 등이 대표적이다. Lord(1980)와 Raju(1988, 1990)의 방법은 처음으로 이분 IRT모형에서 DIF를 평가할 수 있다는 것을 제안한 것으로 이것들은 후에 Cohen, Kim과 Baker(1993)에 의해 다분문항의 IRT 모형으로 확장되었다. Thissen 등(1988)이 내세운 LR검사는 두 집단 간의 문항 응답들 사이의 관찰된 차이들의 유의성을 판단하는 것을 말한다. Flowers, Oshima와 Raju(1999)는 Thissen 등(1988)의 모형과 Raju 등(1995)의 모형은 모두 이분문항과 다분문항에서 DIF검증에 적절하다고 평가하고 있다. 더욱이 Oshima, Raju와 Flowers(1997)에 의하면 Raju 등(1995)의 복수차원의 IRT 모형에도 적절히 이용될 수 있다. 이러한 방법들은 두 모집단들의 문항모수들의 동일성을 평가하는데, 이들과 달리 DFIT에서는 개인의 문항수준에서 진점수의 동일성에 초점을 둔다. 즉, 문항반응함수들이 두 집단에 걸쳐 동일성이 비교되게 된다(Raju et al., 2002).

(2) LR검사 (Likelihood Ratio Test)

앞서 언급한 방법들 중 가장 널리 쓰이는 것은 LR검사로, 문항수준에서 이루어지며 CFA 방법과 같이 최대우도추정(maximum likelihood estimation)이 문항모수들을 추정하는데 사용된다. 이 모수들은 적합함수(fit function)로 알려진 모형적합도 값들에서 기인하는데, IRT에서 적합함수 값은 주어진 모형이 문항모수들을 추정하는데 사용된 최대우도추정 절차의 결과로서 자료와 얼마나 일치하는지 알려주는 역할을 한다(Camilli & Shepard, 1994). LR검사는 기저모형과 비교모형의 일치도를 비교하는 것을 포함한다. 그

식은 다음과 같다.

$$LR_i = \frac{L_C}{L_A} \tag{5}$$

L_C: 기저모형의 우도함수(likelihood function)
L_A: 문항 i의 모수들이 상황에 관계없이 사용되는 비교모형의 우도함수

이 함수의 로그변환이 이루어질 수 있는데, 이는 영가설하의 χ^2 분포의 검정통계량(test statistics)에 기인한다.

$$\chi^2(M) = -2\ln(LR) = -2\ln L_c + 2\ln L_A \tag{6}$$

여기서 M은 기저모형과 비교모형의 추정된 문항모수들의 수의 차이와(즉, 모형들 사이의 자유도) 같은데, LR검사를 이용하기 위해서는 검사의 각 문항들에 대해 χ^2가 계산되어야 하고, 유의미한 χ^2값을 가진 문항들은 DIF를 나타내게 된다(Meade & Lautenschlager, 2004b). 즉, 다른 모수추정치를 사용하는 것은 전체적인 모형적합도를 증가시킨다.

LR 검사를 위해 이전에는 기본적으로 Thissen(1991)의 MULTILOG 프로그램을 여러 번 사용하였으나, Thissen(2001)의 IRTLRDIF 프로그램을 이용하면 좀 더 쉬운 접근이 가능하다. IRTLRDIF 프로그램은 DIF 통계치를 계산하는 여러 단계를 이용하는데, 먼저 모든 문항들에 대한 모수들과 유사한 문항들의 모수들이 시간에 관계없이 동일하다는 제한조건하에 추정된다. 다음으로 유사문항들의 모수추정치들이, 시기에 따라 다르게 측정되는 단일문항의 모수들을 제외하고, 모두 시간에 관계없이 동일하다는 제한조건하에 여러 단계가 이루어진다. 각 문항당 이러한 단계들을 거친 검사의 결과로 각 문항과 관련된 G^2값을 얻을 수 있는데, 이 값들은 문항의 모수들을 자유롭게 하는 것(free)과 관련된 모형 적합도에서의 개선을 나타낸다. 만약 이 값이 유의하면 해당 문항에 대해 DIF가 존재한다고 할 수 있고, 반대로 유의하지 않으면 존재하지 않는다고 판단할 수 있다. 여기서 DIF가 발견된 문항들에 대해서는 IRTLRDIF 프로그램에 의해 추가적인 분석이 수행된다. 물론 이 때도 이전과 같이 모든 문항모수들은 DIF가 있는 문항들

을 제외하고는 시간에 상관없이 동일해야 한다는 것에 제한되어 있다. 이러한 문항들에 대해서는 두 가지 방법이 추가적으로 행해지는데, 먼저 처음에는 모수 b만 시간에 관계없이 동일하다는 상태에서 분석하고, 반대로 두 번째에는 모수 a만 제한되어 있는 상태에서 분석이 행해진다. 이러한 절차들은 각 문항에 대한 DIF의 정확한 원인을 결정할 수 있도록 모수 a와 b의 동일성의 검사를 제공한다. 이러한 LR 검사는, 자료에서 측정 동등성 결여의 원인을 결정하기 위해 모수들을 제한하는 CFA 분석에서의 내재된 카이 자승검사(nested chi-square test)와 상당히 유사하다(Meade, Lautenschlager, & Hecht, 2005).

3. 일반화 가능도이론 (Generalizability Theory, G이론)

1) 이론의 개념과 연구모형

Cronbach, Gleser, Nanda와 Rajaratnam(1972)에 의해 공식적으로 처음 제기된 G이론은 행동측정의 신뢰도에 대한 통계적 이론이다(Cronbach et al., 1972; Shavelson & Webb, 1991). G 이론에서의 신뢰도는 고전검사이론의 신뢰도 개념에 기초한 것이지만, 고전검사이론과는 달리 한 검사점수가 다른 조건 하에서도 일반화될 수 있는지의 여부에 초점을 둔다(Shavelson & Webb, 1991). 예를 들면 검사시점에 상관없이 여러 명의 평가자들의 평가점수가 비슷한지, 혹은 평가자에 상관없이 검사시점마다 비슷한 점수를 얻는지를 분석하는 것이다.

G이론에서는 분산분석을 통해서 검사점수의 일관성에 영향을 미치는 변수들을 조사함으로써 신뢰도계수를 추정한다. 즉, G이론은 오차점수의 다양한 원천을 구별해 내고 각각의 원천이 갖는 상대적인 영향력을 판별해 냄으로써 관찰점수를 진점수와 오차점수로만 구분하는 고전검사이론의 한계를 극복한다(Brennan, 2001; Cronbach et al., 1972).

G이론은 G연구(generalizability study)와 D연구(decision study)로 구분된다. 먼저 G연구는 결과가 얼마나 일반화될 수 있는가에 관심을 갖고, 측정오차에 영향을 주는 다중 오차원을 동시에 분석하며, 분산분석 절차를 이용하여 설계모형에 따른 분산성분을 추정하는 과정이다. 반면 D연구는 주어진 상황에서 가장 효율적인 측정 절차를 결정하기 위한 것으로, 주로 G연구의 결과 산출된 오차원 분산성분의 절대적 또는 상대적 크기에 따라 측정의 조건을 조절함으로써 오차분산을 줄이고 일반화 가능도 계수를 개선할

수 있다. D연구는 측정 기준에 맞고 시간과 비용 측면에서 효율적인 측정 절차를 설계하는데 사용된다.

G이론에서는 고전검사이론과 유사하면서도 구분되는 개념들이 몇 가지 등장한다. 고전검사이론에서의 진점수분산을 전집분산으로, 오차점수분산을 G이론에서 상대오차점수분산과 절대오차점수분산으로 개념화한다. 고전검사이론의 신뢰도 계수와 비슷한 개념으로는 상대오차분산을 이용한 G계수(generalizability coefficient, g-coefficient)가 있다. G계수는 관찰점수분산에 대한 전집점수 분산의 비율로서, 그 식은 다음과 같다.

$$\text{G계수 } \rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{err}^2} \ (\sigma_p^2\text{: 전집점수분산, } \sigma_{err}^2\text{:오차점수분산}) \qquad (7)$$

여기서 오차점수분산은 여러 국면에서 추정된 독립적인 오차점수(error score)분산의 합($\sigma_x^2 = \sigma_p^2 + \sigma_{e1}^2 + \sigma_{e2}^2 + \ldots + \sigma_{ei}^2$)으로 표현된다(단, G계수는 상대오차분산을 오차분산으로 하는 상대결정에 의거한다). 고전검사이론에서 신뢰도 계수와 마찬가지로, 적절한 G계수는 하나의 절대치로 합의된 것은 아니고, 연구의 목적이나 연구자에 따라 이견들이 있다. .9를 절사값(cutoff value)으로 제안한 경우도 있지만(Shavelson & Webb, 1991), 대체적으로 .7 또는 .8 정도가 적절하다고 여겨진다(김성숙, 김양분, 2001; Brennan, 2001). 오차원은 하나일 수도 있고 두 개 이상일 수도 있는데, 예를 들어 오차요인이 하나(문항)인 경우, 피험자(p)의 수행에서 나타나는 피험자 간의 차이는 σ_p^2, 문항(i)에 대한 분산 요인은 σ_i^2, 그리고 잔차(pi,e)에 대한 분산요인은 $\sigma_{\pi,e}^2$이 된다.

G이론은 교차모형, 내재모형, 부분 내재모형, 그리고 국면이 무선효과인 경우와 고정효과인 경우 등 다양한 측정 설계에 적용될 수 있다. 교차모형에서는 한 국면의 모든 조건들이 다른 변수의 모든 조건들과 함께 관찰되는 것으로, 모든 피험자가 모든 문항을 다 응답하였거나, 모든 관찰자가 모든 피험자를 관찰한 경우이다. 측정상황에서 요인이 어느 한 요인 내에 포함되어 있을 때(즉, 내재모형)는 예를 들어 피험자가 각각 다른 문항으로 시험을 본다든가, 면접상황에서 각각 다른 평가자가 면접을 하는 경우, 문항이 피험자에게 내재되어 있을 경우를 의미한다. 무선효과는 국면의 조건들이 무선추출되거나 관찰된 조건들이 무한한 전집에서 표집되어 다른 것으로 얼마든지 대체될 수 있는 경우이다. 반면 고정효과는 일반화가능한 전집이 무한하지 않고 유한전집인 경우를 말하는

데, 예를 들어 만약 관찰대상을 측정하는 관찰자가 변하지 않고 반복하여 같은 표집을 사용한다면 관찰자는 고정국면으로 간주할 수 있다.

2) G이론과 측정동등성: 개념적 접근

G이론은 측정동등성과 관련되어 있으면서도 구분되는 개념이다(Sharma & Weathers, 2003). G이론은 수집된 데이터를 더 큰 전집(larger universe)에 얼마나 일반화할 수 있느냐의 정도를 나타내주는 개념으로, 만일 측정을 일반화할 수 없다면 다른 상황조건 하에서 동일성이 확보될 수 없기 때문에 상이한 집단 간의 측정동등성은 의미가 없게 된다. G이론은 주어진 측정도구가 일반화될 수 있는 정도를 측정하기 위한 지표를 제공할 수 있고, 상이한 상황조건 간의 측정동등성에 대한 추가적인 증거를 제공한다. 또한 한 연구의 결과를 다른 상황에 일반화하는데 있어서 척도 문항과 피실험자들이 얼마나 많이 필요한가를 결정하는 것이 중요하다.

3) G이론을 통한 측정동등성 검증방법

일반화가능도 분석을 위해서는 최소한 효과의 제곱합(sum of square)을 구할 수 있으면 G연구와 D연구 분석에 필요한 통계치를 계산할 수가 있다. 현재 널리 이용되고 있는 프로그램 중 SPSS의 MANOVA 프로그램을 이용해서 제곱합을 구할 수가 있고, BMDP와 SAS의 ANOVA 프로그램을 통해서 일반화가능도 분석을 위한 평균제곱 기대값(expected mean square)을 산출할 수 있다. 그러나 일반 프로그램을 이용하여 G이론을 적용한 분석을 하는 것은 복잡한 수리적 절차 때문에 쉬운 일이 아니다.

GENOVA(Generalized analysis of variance system; Crick & Brennan, 1983) 프로그램의 개발로 G이론에 대한 몇 개의 개념과 간단한 조작방법만을 알면 G연구 분산성분의 추정치뿐만 아니라 상이한 표본의 크기에 따른 D연구 분산 성분인 전집점수 분산, 관찰점수 기대치, 상대오차, 절대오차, G계수, 그리고 준거참조측정의 신뢰도 계수 등을 쉽게 얻을 수 있다.

4. CFA, IRT, G이론의 비교

앞서 상이한 집단 간 비교연구에 있어서 측정동등성을 검증하는 세 가지 기법인 CFA,

IRT, G이론에 대해 앞에서 설명하였다. 이 검증방법들은 각기 서로 다른 특징, 그리고 이점과 한계점이 있는데, 연구자는 이에 대해 인식하고 있어야만 어떤 기법을 이용하여 검증을 할 것인지 결정할 수 있다. 따라서 CFA, IRT 그리고 G이론의 특징 및 장단점을 비교하여 각각 〈표 5〉과 〈표 6〉에 제시하였다.

〈표 5〉 CFA, IRT, 일반화 가능도 이론의 특징

	CFA	IRT	G이론
통계치	$\Delta\chi^2$와 적합도 지수	G^2	G계수
특징	• 선형 모형 • γ변화(gamma change)의 제거, β변화(beta change)가 나타남.	• 로그선형의 가정 • γ변화의 제거가 불가, β변화를 다루는 데 효과적임. • 문항수준의 검사에서는 CFA보다 IRT가 적합함.	• 연구자가 척도를 더 정교하게 하는 것이 불가능할 때 유용하게 사용

자료원: Meade & Lautenschlager, 2004b; Meade, Lautenschlagerm, & Hecht, 2005; Raju et al., 2002 재구성

측정동등성을 검사하기 위한 CFA와 IRT의 접근법은 개념적으로는 유사하나 그 사용에서는 차이를 보인다. 두 방법 모두 관찰된 점수와 잠재 점수들 사이의 관계를 모형화하고, 비동일성이 존재할 때 그 정도를 추정하는 데 사용된다(Raju et al., 2002; Reise, Widaman, & Pugh, 1993). 또한 두 방법 모두 내재모형 비교를 이용할 수 있고 모형적합의 모수 검사들을 제공하기 위해 카이자승검사를 이용할 수 있다. 반면 문항수준의 검사에서는 CFA보다 IRT가 더 적합하다는 점, 그리고 IRT는 CFA와 달리 다른 시점에서의 비슷한 문항들 사이의 공분산이 모형화될 수 없다는 점에서 두 접근법은 차이를 보인다. CFA와 G이론은 모두 척도가 일반화가능하지 않은 정도에 대해 정보를 제공한다. CFA 분석에서는 $\Delta\chi^2$와 적합도 지수가 측정 동등성이 위반되는 정도를 측정해주며, G이론은 상이한 집단 간 척도 항목들의 차이로 인해 발생하는 분산의 정도를 측정해준다. 그러나 CFA는 척도 개발 과정의 초기 단계에서 유용한 반면, G이론은 바람직한 수준의 일반화를 획득하는데 필요한 문항과 피실험자의 수를 결정하는데 대한 가이드라인을 제공하는데 유용한 접근이라 할 수 있다. 또한 IRT는 다분문항으로 처리하여 적용될 때 정보의 손실과 신뢰도 감소를 가져오는 문제가 있지만 G이론은 정보의 손실이

없다.

〈표 6〉 CFA, IRT, G이론의 장단점 비교

검증방법	장점	단점
CFA	• $\Delta\chi^2$를 통해 모형의 적합도의 비교 가능함. • 측정동등성이 성립되지 않을 경우 문제가 되는 항목을 감지할 수 있으므로 척도 개발 과정의 초기 단계에서 유용하며 문제가 되는 항목에 대해 동일성 제약을 해제하여 부분적 측정동등성의 수립이 가능함. • CFA를 이용한 공분산행렬동일성의 검증은 측정동등성 검증에 있어서 효과적임.	• $\Delta\chi^2$를 통계치로 사용할 경우, 표본 크기에 민감함. • 바람직한 수준의 일반화를 획득하는데 필요한 문항과 피실험자의 수에 대한 가이드라인을 제공하지 못함. • 요인 공통성(communality)에 의해 영향을 받아 모든 요인 부하량 값이 작은 경우 모수를 정확하게 추정하지 못하므로 집단 간 차이를 제대로 검증할 수 없음(따라서 적당한 정도의 공통성이 요구).
IRT	• 시간에 따라 모수 b가 변화한다는 것을 보일 수 있으며 시간에 따른 β 변화의 강력한 지표가 될 수 있음. • 문항 모수들이 부분모집단에 의존적이지 않음. • 개개인의 모수들이 검사의 문항들의 집단에 대해 특정하지 않음. • 측정정도가 일정하다고 가정되지 않고, 대신 연구자로 하여금 측정의 조건부 표준오차를 계산하게 함.	• 다분문항으로 처리하여 문항반응이론을 적용할 때 정보의 손실이 발생하며 신뢰도를 감소시킴. • CFA와 달리 다른 시점에서의 비슷한 문항들 사이의 공분산이 모형화될 수 없음.
G 이론	• 바람직한 수준의 일반화를 획득하는데 필요한 문항과 피험자 수를 결정하는데 대한 가이드라인을 제공하므로 효율적인 측정모형이 가능하며 정보의 손실이 없음. • 다른 기법에 비해 적은 컴퓨터 시간과 노력을 요구하며 개념적으로 모형을 이해하기 쉬움. • 다중 오차원을 분석하여 측정 점수 변동요인의 영향력을 비교할 수 있는 분석의 틀을 제공함.	• 상이한 집단 간 척도의 측정동등성이 오차분산의 성분 크기에 의해 주관적으로 평가되므로 통계적 엄밀성이 감소함. • 척도 개발 과정이나 부분적 측정동등성을 수립하는 절차에서 진단 정보(diagnostic information)를 제공하지 않음. • 무선표집에 대한 가정이 만족되지 않으면 결과가 과대추정됨.

자료원: 김성숙 & 김양분, 2001; 이규민, 2003; Lee, 2000a, 2000b; Meade & Lautenschlager, 2004b; Rouse, Finger, & Butcher, 1999; Sharma & Weathers, 2003; Yen, 1993 재구성

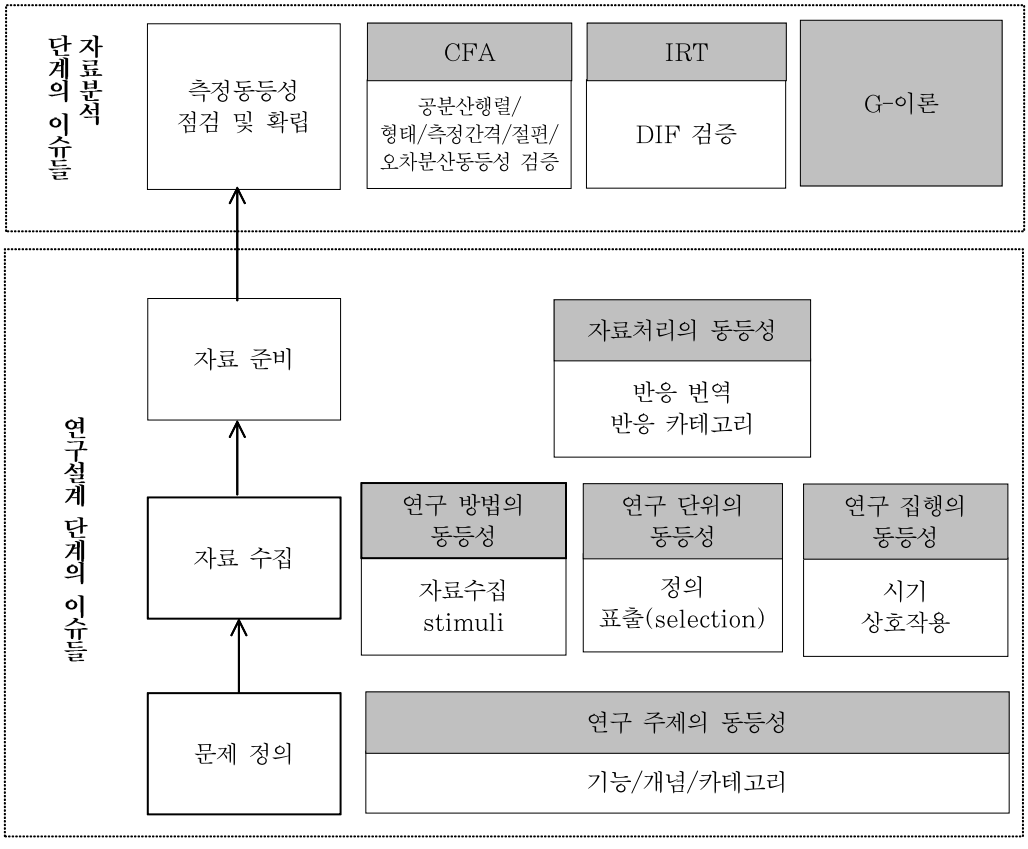
이상적으로는 측정동등성을 조사할 때 CFA와 IRT, 일반화 가능도 이론을 모두 이용하는 것이 좋다. 먼저 IRT를 이용하여 문항수준에서 각 척도 혹은 하부척도 내에서 측정동등성을 측정할 수 있고, 이러한 조건들을 만족하는 문항들에 대해 CFA를 이용하여 개인의 척도와 더 복잡한 모형에서 측정동등성을 측정할 수 있다(Meade & Lautehschlager, 2004b). G이론은 G계수를 제공함으로써 연구자가 척도를 더 정교하게 하는 것이 불가능할 때 유용하게 사용될 수 있다. G계수가 크면 그 척도가 상이한 집단 간 비교연구에 사용될 수 있고, G계수가 작으면 문항 중 일부가 집단 간 동일한 의미를 지니지 않는다는 것을 나타내 줄 수도 있다. 이 경우 CFA를 사용하여 측정 비동일성을 감지하고 문제문항을 제거할 수 있다. 결론적으로 CFA, IRT 그리고 G이론은 측정동등성을 유지하기 위해 상호 보완적으로 사용될 수 있다.

V. 측정동등성 문제의 극복방안과 제언

1. 연구단계별 측정동등성 문제의 극복방안

지금까지 논의한 측정동등성의 문제를 극복하기 위해서 연구단계별로 무엇을 어떻게 해야 하는지를 〈그림 3〉에 제시하였다. 〈그림 3〉에서는 연구설계 단계와 자료분석의 단계를 구분하여 측정동등성의 사전적, 사후적 검증 방법을 논의하고자 한다. 사전적 측정동등성 확보를 위한 연구설계 단계에서는 통계적인 검증에 앞서 연구자가 측정동등성을 획득하기 위해 취할 수 있는 방법을 제시한다. 사후적 검증인 자료분석 단계에서는 앞서 논의했던 여러 통계기법들을 사용하여 측정동등성을 검증하는 단계이다.

사전적으로 연구설계 단계에서는 (1) 연구주제(research topics) (2) 연구방법(research methods) (3) 연구단위(research units) (4) 연구집행(research administration) (5) 자료처리(data handling)의 주요한 다섯 가지 영역에서 동일성을 고려해야 한다. 이 중 연구주제의 동일성은 연구방법론적 관점에서 볼 때 나머지 영역들에 비해 가시성이 덜하지만 구성체의 측정척도를 개발하는 것과 관련이 있기 때문에 매우 중요한 영역이라 할 수 있다(Albaum & Baker, 2005). 문제 정의 단계에서, 연구주제의 동일성



자료원: Holzmuller & Salzberger, 1999; Salzberger, Sinkovics, & Schlegelmilch, 1999; 수정인용

〈그림 3〉 연구단계별 측정동등성 검증방법

은 상이한 집단 간의 양적 비교에 있어서 필요한 최소의 기준이라 할 수 있는데, 만일 연구하고자 하는 구성체가 다른 문화에서 다른 기능을 가지고 있거나 개념적으로 동등하지 않다면 이 구성체 개념에 대한 연구자체가 의미 없게 되기 때문이다.

문제정의가 제대로 이루어지면, 그 다음은 상대적으로 기술적 분야인 자료수집 및 처리단계에서 동일성을 점검해야 한다. 우선 자료수집 단계에서 동일성을 고려해야 한다. 예를 들어, 연구방법의 동일성은 문화 특수적인 반응 왜곡을 최소화하기 위해서 적절한 단계들을 취해야 할 뿐만 아니라, 문화적으로 적절한 자료수집 기술을 선택할 것을 요구한다(Casley & Lury, 1981; Kracmar, 1971; Salzberger, Sinkovics, & Schlegelmilch,

1999). 연구단위의 동일성을 확보하는 것은 연구되는 집단들의 전체 모집단을 정의하고 표본추출을 하는 방식에 영향을 주게 되는데, 여기서 근본적인 문제가 되는 것은 문화 내적으로 대표성을 갖는 표본과 이문화 간 비교 가능한 표본의 상쇄(trade-off) 문제이다. 또한, 연구집행의 동일성은 수집된 자료의 질을 손상시키는 구체적인 변수들(예: 시간 관련 변수, 정치적 변수 등)이 양 문화에서 모두 없다는 것을 확인하는 것이다. 자료수집 이후에 자료준비 단계에서는 데이터가 동등하게 다뤄지고 있는지(예: 동등한 반응 번역 등)에 주의를 기울여야 한다. 이와 같은 동일성의 다양한 영역들은 서로 긴밀히 연결되어 있어 비동일성이 나타날 경우 그것이 어디에서 기인하는 것인지 분명히 아는 것이 쉽지 않다(Salzberger, Sinkovics & Schlegelmilch, 1999).

상이한 집단 간의 양적연구는 최종적으로 자료의 동일성을 필요로 한다. 자료동일성이 획득되었는지의 여부는 언제나 실증적 이슈가 되기 때문이다. 이를 위해 본문에서 논의된 바와 같이 사후적으로 측정동등성을 검증하는데 적합한 CFA, IRT, G이론과 같은 통계적 절차들이 수행된다. 만일 만약 이러한 절차를 통해 자료동일성이 유지되지 않았다면, 연구는 집단 간 의미 없는 양적 비교를 수행하는 것보다는 연구의 목표를 질적 해석에 제한하는 것이 바람직하다.

2. 측정동등성 연구의 이슈와 미래 연구를 위한 제언

측정동등성은 1930년대부터 그 논의가 시작되었음에도 아직까지 그 용어에 대한 합의나 구체적인 검증방법에 대한 합의가 완전히 되지 않아 지속적으로 논쟁이 이루어지고 있는 상태이다. 우선 측정동등성 검증의 대표적 방법인 CFA에 대해서는 검증의 순서 및 부분적 측정동등성이 논쟁이 되고 있다. 이에 대한 논의와 관련하여, 첫 번째로 연구자들마다 어느 검증방법을 사용해야 하며 어느 순서로 검증을 할지에 대하여 서로 다른 의견이 제시되고 있다. Vandenberg와 Lance(2000)에 따르면 가장 많이 사용되는 측정동등성의 검증방법은 측정간격동일성, 형태동일성, 오차분산동일성 순이며, 순서에 있어서는 공분산행렬동일성을 가장 먼저 검증하고 형태적동일성을 기저모형으로 검증한다는 것에 대해서 대체적인 동의가 있다고 한다. 하지만 그 이후의 순서에 있어서는 연구자들 간에 서로 다른 견해를 보이고 있다. 즉, 측정동등성의 검증방법에 대해 앞서 언급한 방법을 모두 사용해야 하는지, 아니면 그 중 몇 개만 선택적으로 사용해야 하는지에

대한 합의는 아직 이루어지지 않고 있으므로, 이에 대한 종합적 연구가 필요하다.

두 번째로 CFA에서 부분적 측정동등성에 대해서는 집단 간 측정의 차이를 어느 정도까지 인정해야 할 것인가의 문제가 있다. 즉, 모수가 어느 정도까지 서로 다르도록 허용할 것인가를 결정하는 문제에 대한 합의가 없다는 것이다. Steenkamp와 Baumgartner (1998)는 어느 모수가 동일성 제약을 받지 않아도 되는가에 대한 고려할 때, 수정지수(modification indices)와 예상모수변화(expected parameter changes)를 이용할 수 있다고 했다. 수정지수 값이 매우 유의할 때 동일성 제약을 풀도록 권장되며, 일반적으로 모형의 수정은 최소로 하여 모형적합성의 심각한 문제에 대해서만 모형의 수정이 이루어져야 한다는 것이다. 반면 Hair 등(2006)에서는 부분적 측정간격동일성이 유지되기 위해서는 집단 간 부하량의 추정치가 적어도 두 요소 이상에서 같아야 하며, 마찬가지로 적어도 두 영절편(zero-intercept)이 같을 때 부분적 절편동일성이 유지된다고 한다. 앞으로 이에 대한 연구와 합의가 이루어질 것으로 기대된다.

IRT에 대해서는 상대측정의 결과에 대한 이슈가 존재한다. 이 문제는 종단연구와 같이 해마다 혹은 분기마다 반복적으로 이루어지는 조사에서 빈번히 발생할 수 있는 문제이다. 이와 같이 다른 측정을 경험하는 개인들을 비교하기 위한 방법들은 검사의 동등화 과정(test-equating procedures, Doran & Holland, 2000)의 주제이기도 하다. 최근에는 IRT를 이용하여 같은 개념을 다른 척도들을 이용하여 측정했을 때 이들을 같은 척도로 가능하도록 변환하는 쪽으로 관심이 쏠리고 있는데(Choi & McCall, 2002), 이것들은 기본적으로 각 측정에 DIF가 없다는 것을 가정한다는 점에서 한계가 있다.

측정동등성의 구체적 검증방법에 대하여 앞의 논의사항들에 대해 보다 많은 관심을 가지고 연구를 진행하여야 할 것이다. 이에 더하여 미래의 연구과제로 앞에서 제시한 세 가지 검증방법들에 대한 통합적인 실증연구를 제안한다. 측정동등성의 검증에 있어서 CFA, IRT, G이론에 대한 연구가 있어왔지만, 대부분 CFA를 중심적으로 연구하였으며 CFA와 IRT의 비교(예: Meade & Lautenschlager, 2004b; Meade, Lautenschlager, & Hecht, 2005), CFA와 G이론의 비교(예: Sharma & Weathers, 2003)에 관한 연구는 진행되어왔으나, 이 세 가지 이론에 대한 통합적인 연구가 부족했다. 세 가지 기법을 사용한 종합적 실증분석을 통하여 이들이 각기 어느 정도 측정비동일성을 감지해 내는지에 대해 그 정도의 차이를 비교 분석하는 것이 필요하다.

Ⅵ. 결 론

본 연구는 측정동등성의 의미, 중요성, 그리고 검증방법을 살펴봄으로써 궁극적으로 연구에서 측정동등성의 중요성을 주장하고자 했다. 과학적 추론을 위해서는 측정의 동일성에 대한 증거를 확보하는 것이 중요하다(Horn & McArdle, 1992). 경영학 연구에 있어서 연구자들은 측정동등성에 대한 증거를 종종 제시하지 않고 연구를 진행하곤 한다. 즉, 대부분의 국가, 문화, 시간에 대한 비교연구에서 집단 간 측정동등성을 가정하지만 측정동등성에 대한 관심이 부족했으며, 국내 연구에 있어서 측정동등성을 검증한 논문은 찾아보기 힘들었다. 이와 같이 측정동등성에 대한 증거가 제시되지 않을 경우, 외적타당성, 예측타당성, 크로스확인에 위협요소가 될 수 있다. 타당성이 위협받게 되면 연구 표본에서 도출된 이론을 일반화시키는 것이 어렵게 되며, 이는 학문적 발전에 있어서의 공헌도 있는 이론의 생성을 저해할 수 있다. 따라서 인사조직, 마케팅, 국제경영 등, 연구에 있어 측정이 중요한 역할을 차지하는 분야를 연구하는 학자들은 측정동등성의 중요성을 인식하고 연구설계 단계에서 측정동등성의 검증을 포함하여 측정동등성이 확보된 후 연구를 진행시켜야 할 것이다.

측정동등성은 앞에서 언급한 바와 같이 CFA, IRT, 그리고 G이론을 통해 검증할 수 있다. 먼저 CFA를 기반으로 하여 집단 간 개념의 동질성, 측정변수와 이론적 변수 간의 동등한 관계, 오차분산의 동일성에 대해 검증할 수 있다. 또한 IRT를 바탕으로 하여 DIF 검증 중에서 가장 많이 사용되는 LR검사는 집단 간 문항 모수들의 동일성을 평가한다. 마지막으로 G 이론은 집단 간 척도 문항들의 차이로 발생하는 분산의 정도를 측정함으로써 CFA와 IRT에 추가적으로 상이한 상황조건 간의 측정동등성에 대한 증거를 제시할 수 있다. 앞에서 언급했듯 측정동등성 검증의 세 가지 이론들은 각기 장단점을 지니고 있다. 따라서 세 가지를 모두 사용하여 측정의 동일성을 확인하는 것이 가장 바람직하겠으나 현실적으로 불가능하다면 연구자의 측정이론에 따라 적절한 방법을 결정해야 할 것이다. 이 밖에도 측정동등성 검증에 있어서 여러 가지 이슈들이 있다. CFA 검증에서는 어떤 순서로 검증할지와 어떤 종류의 검증을 사용해야 할지, 부분적 동일성을 수용할 것인지에 대한 논란이 있는데, 이에 대해 가장 좋은 방법(one-best way)이 존재한다기보다는 각자의 연구목적에 맞는 방법을 선택해야 할 것이다. 또한 IRT에 있어서도 상대측정의 결과를 어떻게 비교해야 할 것인가의 문제가 있다.

요약하면, 측정동등성의 검증을 실행함에 있어서 연구자는 CFA, IRT, G이론의 이점과 한계점에 대해 인식해야 하며, 그 밖의 측정동등성을 둘러싼 이슈들을 고려하여 그 검증방법을 결정하고, 측정동등성을 확인해야 한다. 만약 측정동등성이 검증되지 않는다면 그 측정을 이용한 연구는 타당성이 의심될 것이므로 측정의 수정 등을 통해 그 동일성을 확보해야 할 것이다. 향후 경영학 연구자들이 본 연구를 바탕으로 측정동등성 문제를 인식하여 해당 분야 연구를 수행함에 있어 그 타당성을 한층 높이며, 나아가 관련된 방법론에 대한 논의가 더욱 활성화 되는 계기가 되기를 기대한다.

참 고 문 헌

김성숙 & 김양분. 2001.「일반화가능도 이론」, 서울: 교육과학사.

강정애. 1997. 조직문화적 특성에 따른 조직성과에 관한 연구: Rosseau의 조직문화 모형을 중심으로,「경영학연구」, 26: 513-530.

강혜련. 1998. 리더십과 조직적응: 남녀관리자의 비교연구,「인사·조직연구」, 6(2): 81-123.

권순식 & 김상진. 2005. 근로자의 공정성 지각과 조직시민행위: 고용형태가 미치는 영향에 대한 탐색적 연구,「인사·조직연구」, 13: 1-34.

김경수. 1998. 개인과 조직의 일치: 인턴사원제도가 조직구성원의 사회화와 작업결과에 미치는 효과,「경영학연구」, 27: 1003-1024.

김규남 & 신민수. 2001. 한국, 인도네시아, 중국종업원 조직몰입특성에서 개인적 가치성향의 매개효과에 대한 연구,「경영학연구」, 30: 877-904.

김영조. 2000, 조직문화와 조직성과의 변화에 관한 종단적 연구,「인사·조직연구」, 8(2): 111-134.

김정구, 김태웅 & 박승배. 2003. 온라인 게이머의 라이프스타일에 관한 탐색적 연구: 비게이머와의 비교 및 게이머 세분시장별 비교.「경영학연구」, 32: 1741-1770.

김정원, 김태형 & 권중생. 2004. 고용형태 및 고용관계특성과 직무몰입간의 관계: 고용형태의 직접효과와 조절효과검증,「인사관리연구」, 28(6): 23-50.

김학수. 1997. 종업원의 노력회피 성향과 그 원인에 관한 연구,「경영학연구」, 26(1):

37-65.

박원우, 김미숙, 정상명 & 허규만. 2007. 동일방법편의(Common Method Bias)의 원인과 해결방안, 「인사·조직연구」, 15(1): 89-133.

박철 & 이태민. 2006, 온라인 구전효과에 영향을 미치는 요인에 관한 비교문화적 실증연구, 「경영학연구」, 35(6): 1617-1647.

서문식 & 김상희. 2004. 인터넷 쇼핑에 있어 성별에 따른 감정적 반응의 종단적 연구, 「경영학연구」, 33(3): 703-739.

성태제. 2001. 「문항반응이론의 이해와 적용」, 서울: 교육과학사.

엄명용 & 김태웅. 2006. 성별 차이를 중심으로 본 이러닝 만족도 영향요인에 관한 연구, 「경영학연구」, 35: 51-80.

이규민. 2003. 단위검사 개념의 적용: 일반화가능도 이론을 중심으로, 「교육평가연구」, 16(1): 53-70

이덕로 & 서도원. 1998, 한국기업의 경영특성에 관한 종단적 연구, 「경영학연구」, 27(4): 911-936.

이순묵. 1992. 심리검사 제작의 이론과 실제: 검사의 동등화, 한국심리학회 1992년 제10회 동계연수회 발표 논문집, 서울.

이지우 & 김종우. 2002. 고용형태에 따른 직무특성과 조직몰입의 관계, 「인사·조직연구」, 10(1): 1-26.

이태식. 2005. 시간지향성 및 시간인식에 관한 한불간 비교문화적 실증 연구, 「인사관리연구」, 29(3): 175-197.

장동운. 2003, 이문화권에서의 개인간 갈등관리의 성별간 비교연구, 「인사관리연구」, 27(2): 61-82.

장은주 & 박경규. 2005. 성별에 따른 개인특성 및 사회적 자본과 주관적 경력성공과의 관계, 「경영학연구」, 34(1): 141-168.

조용래 & 김정호. 2002. 한국판 Beck Depression Inventory의 확인적 요인분석: 대학생과 임상표본 간 구조 및 측정동일성 검증, 「한국심리학회지: 임상」, 21: 843-857.

황호종. 2004. 국가 간 문화적 차이에 측정 동등성을 유지키 위한 방법론에 관한 탐색적 연구, 「경영교육논총」, 35: 127-140.

Albaum, G., & Baker. K. G. 2005. The imposed etic in survey research: fact or fallacy?, *Proceedings of the 2005 International Business and Economy Conference*, Honolulu, HI: January.

Ambrose, M. L. & Cropanzano. R. 2003. A longitudinal analysis of organizational fairness: An examination of reactions to tenure and promotion decisions, *Journal of Applied Psychology*, 88: 266-275.

Avery, D. R. 2003. Reactions to diversity in recruitment advertising: Are differences black and white?, *Journal of Applied Psychology*, 88: 672-679.

Azevedo, A., Drost. E. A., & Mullen. M. 2002. Individualism and collectivism: Toward a strategy for testing measurement equivalence across culturally diverse groups, *Cross Cultural Management,* 9(1): 19-29.

Azocar, F., Arean. P., Miranda. J., & Munoz. R. F. 2001. Differential item functioning in a Spanish translation of the Beck Depression Inventory, *Journal of Clinical Psychology*, 57(3): 355-365.

Bagozzi, R. P., Verbeke, W., & Gavino. J. C. Jr. 2003. Culture moderates the self-regulation of shame and its effects on performance: The case of salespersons in the Netherlands and the Philippines, *Journal of Applied Psychology*, 88: 219-233.

Bauer, T. N., Erdogan, B. R., Liden, C., & Wayne, S. J. 2006. A longitudinal study of the. moderating role of extraversion: LMX, performance, and turnover during new executive. Development, *Journal of Applied Psychology,* 91: 298-310.

Begley, T., & Lee, C. 2005. The role of negative affectivity in pay-at-risk reactions: A longitudinal study, *Journal of Applied Psychology*, 90: 382-388.

Berry, J. W. 1980. Introduction to methodology, In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (vol. 2): 1-28. Boston, MA: Allyn and Bacoon.

Bollen, K. A. 1989. ***Structural equations with latent variables***, NY: John Wiley.

Borsboom, D., Mellenbergh, G. J., & Heerden, J. V. 2002. Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias, ***Applied Psychological Measurement***, 26(4): 433-450.

Brennan, R. L. 2001. ***Generalizability Theory***, NY: Springer-Verlag.

Byrne, B. M. 1994. Testing for the factorial validity, replication, and invariance of a measurement instrument: A paradigmatic application based on the Maslach burnout inventory, ***Multivariate Behavioral Research,*** 29: 289-311.

Byrne, B. M., Shavelson, R. J., & Muthen, B. 1989. Testing for the factor covariance and mean structures: The issue of partial measurement invariance, ***Psychological Bulletin***, 105: 456-466.

Cable, D. M. & DeRue, D. S. 2002. The convergent and discriminant validity of subjective fit perception, ***Journal of Applied Psychology***, 87: 875-884.

Camilli, G. & Shepard, L. A. 1994. ***Methods for identifying biased test items***, Thousand Oaks, CA: Sage.

Casley, D. J. & Lury, D. A. 1981. ***Data collection in developing countries***, Oxford: Clrendon Press.

Cella, D., Hernandez, L., Bonomi, A. E., Corona, M., Vaquero, M., Shiomoto, G., & Baez, L. 1998. Spanish language translation and initial validation of the functional assessment of cancer therapy quality of life instrument. ***Medical Care***, 36(9): 1407-1418.

Cella D., Lloyd, S. R., & Wright, B. 1996. Cross-cultural instrument equating: current research and future directions. In B. Spilker (Ed.), ***Quality of Life and Pharmacoeconomics in Clinical Trials*** (2nd ed.): 707-715, Philadelphia, PA: Lippincott-Raven.

Chan, A. & Rossiter, J. 2003. Measurement issues in cross-cultural values research, *Paper presented at the the ANZMAC 2003*, Adelaide: Australia.

Chatman, J. A. & O'Reilly, C. A. 2004. Asymmetric reactions to work group sex diversity among men and women, *Academy of Management Journal,* 47: 193-208.

Chen, G. 2005. Newcomer adaption in teams: Multilevel antecedents and outcomes, *Academy of Management Journal*, 48: 101-116.

Chernyshenko, O. S., Stark, S., Chan, K-Y., Drasgow, F., & Williams, B. 2001. Fitting item response theory models to two personality inventories: Issues and insights, *Multivariate Behavioral Research,* 36(4): 523-562.

Cheung, G. W. & Rensvold, R. B. 1999. Testing factorial invariance across groups: A reconceptualization and proposed new method, *Journal of Management*, 25: 1-27.

Choi, S. W. & McCall, M. 2002. Linking bilingual mathematics assessments: A monolingual IRT approach, In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation:* 317-338. Mahwah, NJ: Lawren Erlbaum Associates.

Cohen, A. S., Kim, S. H., & Baker, F. B. 1993. Detection of differential item functioning in the graded response model, *Applied Psychological Measurement*, 17: 335-350.

Cole, M. S., Bedeian, A. G., & Field, H. S. 2006. The measurement equivalence of web-based and paper-and-pencil measures of transformational leadership: A multinational test, *Organizational Research Methods*, 9(3): 339-368.

Crick, J. E. & Brennan, R. L. 1983. *Manual for GENOVA: A Generalized Analysis of Variance System*, Iowa City, IA: The American College

Testing Program.

Crocker, L. & Algina, J. 1986. *Introduction to classical and modern test theory*, Orlando, FL: Holt, Rinehart and Winston.

Cronbach, L.J., Gleser, G. C., Nanda, H., & Rajaratnam, N. 1972. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* NY: John Wiley.

Cullen, J. B. & Parboteeah, K. P. 2004. Cross-national differences in managers' willingness to justify ethically suspect behaviors: A test of institutional anomie theory, *Academy of Management Journal,* 47: 411-421.

De Jonge, J. & Dormann, C. 2006. Stressors, resources, and strain at work: A longitudinal test of the triple-match principle, *Journal of Applied Psychology*, 91: 1359-74.

Donovan, J. J. & Williams, K. J. 2003. Missing the mark: Effects of time and causal attributions on goal revision in response to goal-performance discrepancies, *Journal of Applied Psychology*, 88: 379-390.

Doran, N. J., & Holland, P. W. 2000. Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37: 281-306.

Douglas, S. P. & Craig, C. S. 1983. *International Marketing Research*. Englewood Cliffs, NJ: Prentice-Hall.

Drasgow, F. 1984. Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues, *Psychological Bulletin,* 95: 134-135.

Drasgow, F. 1987. Study of measurement bias of two standardized psychological test, *Journal of Applied Psychology*, 72: 19-29.

Eddleston, K. A., Veiga, J. F., & Powell, G. N. 2006. Explaining sex differences in managerial career satisfier preferences: The role of gender self-schema, *Journal of Applied Psychology*, 91: 437-445.

Epitropaki, O. & Martin, R. 2005. From ideal to real: A longitudinal study

of the role of implicit leadership theories on leader-member exchanges and employee outcomes, *Journal of Applied Psychology,* 90: 659-676.

Flaherty, J. A. 1987. Appropriate and inappropriate research methodologies for Hispanic mental health. In M. Gaviria (Ed.), *Health and behavior: Research agenda for Hispanics:* 177-186, Chicago: University of Illinois Press.

Flowers, C. P., Oshima T. C., & Raju, N. S. 1999. A description and demonstration of the polytomous-DFIT framework, *Applied Psychological Measurement*, 23: 309-326.

Fritz, C. & Sonnetag, S. 2006. Recovery, well-being, and performance-related outcomes: The role of workload and vacation experiences, *Journal of Applied Psychology*, 91: 936-945.

Fullagar, C. J., Gallagher, D. G., Clark, P. F., & Carroll, A. E. 2004. Union commitment and participation: A 10-year longitudinal study, *Journal of Applied Psychology*, 89, 730-737.

Fuller, J. A., Fisher, G. G., Stanton, J. M., Spitzmuller, C. , Russell, S. S., & Smith, P. C. 2003. A lengthy look at the daily grind: Time series analysis of events, mood, stress, and satisfaction, *Journal of Applied Psychology*, 88, 1019-1033.

Gaski, J. F. & Etzel, M. J. 1986. The index of consumer sentiment toward marketing, *Journal of Marketing*, 56: 71-81.

Gelfand, M. J., Higgins, M., Nishii, L. H., Raver, J. L., Dominguez, A., Murakami, F., & Yamaguchi, S. 2002. Culture and egocentrism perceptions of fairness in conflict and negotiation, *Journal of Applied Psychology*, 87, 833-845.

Gong, Y. & Fan, J. 2006. Longitudinal examination of the role of goal orientation in cross-cultural adjustment, *Journal of Applied Psychology*, 91: 176-184.

Green, R. T. & White, P. D. 1976. Methodological consideratoins in cross-

national consumer research, *Journal of International Business Studies*, 7(2): 81-86.

Hair, J. F. Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. 2006. *Multivariate data analysis* (6th ed.), Upper Saddle River, NJ: Pearson Education.

Herdman, M., Fox-Rushby, J., & Badia, X. 1997. 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. *Quality of Life Research,* 6(3): 237-247.

Holzmuller, H. H. & Salzberger, T. 1999. June, Solving the Puzzle of Equivalence in Cross-National Marketing Research: Concepts, Assessments, and Research Designs. *Paper presented at the 9th Biennial World Marketing Congress*: 23-26, Malta.

Horn, J. L. & McArdle, J. J. 1992. A practical and theoretical guide to measurement invariance in aging research, *Experimental Aging Research*, 18: 117-144.

Hui, C. H. & Triandis, H. C. 1985. Measurement in cross-cultural psychology: A review and comparison of strategies, *Journal of Cross-Cultural Psychology,* 16: 131-152.

Jackson, P., Wall, T., Martin, R., & Davids, K. 1993. New measures of job control, cognitive demand, and production responsibility, *Journal of Applied Psychology*, 78: 753-762.

Johnson, M. D., Morgeson, F. P., Ilgen, D. R., Meyer, C., & Lloyd, J. R. 2006. Multiple professional identities: Examining differences in identification across work-related targets, *Journal of Applied Psychology*, 91: 498-506.

Joreskog, K. G. 1971. Simultaneous factor analysis in several populations, *Psychometrika*, 36: 409-426.

Keller, R. T. 2006. Transformational leadership, initiating structure, and substitutes for leadership: A longitudinal study of research and development

project team performance, *Journal of Applied Psychology*, 91: 202-210.

Kracmar, J. Z. 1971. *Marketing research in developing countries*, NY: Praeger Publishers.

Labouvie, E., & Ruetsch, C. 1995. Testing for equivalence of measurement scales: Simple structure and metric invariance revisited, *Multivariate Behavioral Research,* 30: 63-76.

Lam, S. S. K., Chen, X., & Schaubroeck, J. 2002. Participative decision making and employee performance in different cultures: The moderating effects of allocentrism/idiocentrism and efficacy, *Academy of Management Journal*, 45, 905-914.

Lastovicka, J. L. 1982. On the validation of lifestyle traits: A review and illustration, *Journal of Marketing Research*, 19: 126-138.

Lee, G. 2000a. A comparison of methods of estimating conditional standard errors of measurement for testlet-based test scores using simulation techniques, *Journal of Educational Measurement*, 37: 91-112.

Lee, G. 2000b. Estimating conditional standard errors of measurement for tests composed of testlets, *Applied Measurement in Education,* 13: 161-180.

Lester, S. W. & Meglino, B. M. 2002. The antecedents and consequences of group potency: A longitudinal investigation of newly formed work groups, *Academy of Management Journal*, 45: 352-368.

Liu, C., Borg, I., & Spector, P. E. 2004. Measurement equivalence of the german job satisfaction survey used in a multinational organization: Implications of Schwartz's culture model, *Journal of Applied Psychology*, 89: 1070-1082.

Lord, F. M. 1980. *Applications of item response theory to practical testing problems*, Hillsdale, NJ: Lawrence Erlbaum.

Marsh, H. W. 1993. The multidimensional structure of academic self-concept:

Invariance over gender and age, *American Educational Research Journal*, 30(4): 841-860.

Meade, A. W. & Lautenchlager, G. J. 2004a. A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/Invariance, *Structural Equation Modeling*, 11(1): 60-72.

Meade, A. W. & Lautenschlager, G. L. 2004b. A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance, *Organizational Research Methods*, 7: 361-388.

Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. 2005. Establishing measurement equivalence and invariance in longitudinal data with item response?theory, *International Journal of Testing*, 5(3): 279-300.

Mellenbergh, G. J. 1989. Item bias and item response theory, *Psychological Bulletin*, 115: 300-307.

Meredith, W. 1964a. Notes on factorial invariance, *Psychometrika*, 29: 177-185.

Meredith, W. 1964b. Rotation to achieve factorial invariance, *Psychometrika*, 29: 187-206.

Meredith, W. 1993. Measurement invariance, factor analysis and factorial invariance, *Psychometrika*, 58: 525-543.

Millsap, R. E. & Everson, H. T. 1993. Methodology review: Statistical approaches for assessing bias, *Applied Psychological Measurement*, 17: 297-334.

Mullen, M. R. 1995. Diagnosing measurement equivalence in cross-national research, *Journal of International Business Studies*, 3: 573-596.

Oshima, T. C., Raju, N. S., & Flowers, C. 1997. Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and test, *Journal of Educational Measurement,* 34: 253-272.

Perrewe, P. L., Zellars, K. L, Ferris, G. R., Rossi, A. M., Kacmar, C. J., &

Ralston, D. A. 2004. Neutralizing job stressors: Political skill as an antidote to the dysfunctional consequences of role conflict, *Academy of Management Journal*, 47: 141-152.

Polzer, J. T., Crisp, C. B., Harvenpaa, C. L., & Kim, J. W. 2006. Extending the faultline model to geographically dispersed teams; How colocated subgroups can impair group functioning, *Academy of Management Journal*, 49: 679-692.

Porath, C. L. & Bateman, T. S. 2006. Self-regulation: From goal orientation to job performance. *Journal of Applied Psychology*, 91(1): 185-192.

Raju, N. S. 1988. The area between two item characteristic curves, *Psychometrika*, 53: 495-502.

Raju, N. S. 1990. Determining the significance of estimated significance of estimated signed and unsigned areas between two item response functions, *Applied Psychological Measurement*, 14: 197-207.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. 2002. Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory, *Journal of Applied Psychology*, 87: 517-529.

Raju, N. S., Van der Linden, W. J., & Fleer, P. F. 1995. IRT-based internal measures of differential functioning of items and tests, *Applied Psychological Measurement*, 19: 353-368.

Raykov, T. 2004. Behavioral scale reliability and measurement invariance evaluation using latent variable modeling, *Behavior Therapy*, 35: 299-331.

Reeve, C. L. & Lam, H. 2005. The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes, *Intelligence*, 33: 535-549.

Reise, S. P., Widaman, K. F., & Pugh, R. H. 1993. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance, *Psychological Bulletin*, 114: 552-566.

Reise, S. P. & Henson, J. M. 2003. A discussion of modern versus traditional psychometrics as applied to personality assessment scales, *Journal of Personality Assessment*, 82: 93-103.

Robie, C., Zickar, M. J., & Schmit, M. J. 2001. Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales, *Human Performance*, 14: 187-207.

Rouse, S. V., Finger, M. S., & Butcher, J. N. 1999. Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 Psy-5 scales, *Journal of Personality Assessment*, 72: 282-307.

Saks, S. M. & Ashforth, B. E. 2002. Is job search related to employment quality? It all depends on the fit, *Journal of Applied Psychology*, 87: 646-654.

Salzberger, T., Sinkovics, R. R., & Schlegelmilch, B. B. 1999. Data equivalence in cross-cultural research: A comparison of classical test theory and latent trait theory based approaches, *Australasian marketing journal*, 7(2): 23-38.

Samejima, F. 1969. Estimation of latent ability using a response pattern of graded scores, *Psychometrika Monograph Supplement*, 34: 100-114.

Schmit, M. J. & Ryan, A. M. 1993. The big five in personnel selection; Factor structure in applicant and nonapplicant populations, *Journal of Applied Psychology*, 78: 966-974.

Schneider, B., Hanges, P. J., Smith, D, B., & Salvaggio, A. N. 2003. Which comes first: Employee attitudes or organizational financial and market performance?, *Journal of Applied Psychology*, 88(5): 836-851

Shaffer, M. A., Harrison, D. A., Gregersen, H., Black, J. S., & Ferzandi, L. A. 2006, You can take it with you: Individual differences and expatriate effectiveness, *Journal of Applied Psychology*, 91(1), 109-125.

Sharma, S. & Weathers, D. 2003. Assessing generalizabilty of scales used in cross-national research, *International Journal of Research in Marketing*,

20(3): 287-295.

Shavelson, R. J. & Webb, N. M. 1991. *Generalizability Theory: A Primer*, Newbury Park, CA: Sage.

Shaw, J. D., Duffy, M .K., Lockhart, D. E., Mitra, A., & Bowler, M. 2003. Reactions to merit pay increases: A longitudinal test of a signal sensitivity perspective, *Journal of Applied Psychology*, 88: 538-544.

Shultz, K. S. & Whitney, D. J. 2005. *Measurement theory in action: Case studies and exercises*, Thousand Oak, CA: Sage.

Simmering, M. J., Colquitt, J. A., Noe, R. A., & Porter, C. O. 2003. Conscientiousness, autonomy fit, and developmnet: A longitudinal study, *Journal of Applied Psychology,* 88: 954-963.

Singh, J. 1995. Measurement issues in cross-national research. *Journal of International Business Studies*, 26(3): 597-619.

Simpson, P. A., & Stroh, L. K. 2004. Gender differences: Emotional expression and feelings of personal inauthenticity, *Journal of Applied Psychology*, 89: 715-721.

Smith, T. W. 2004. Developing and evaluating cross-national survey instruments. In S. Presser, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, J. M. Rothgeb, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires:* 431-452, River Street, NJ: Wiley.

Smither, J. W. & Walker, A. G. 2004. Are the characteristics of narrative comments related to improvement in multirater feedback ratings over time?, *Journal of Applied Psychology*, 89, 575-581.

Spearman, C. 1904. The proof and measurement of association between two things, *American Journal of Psychology*, 15: 72-101.

Spicer, A., Dunfee, T. W., & Bailey, W. J. 2004. Does national context matter in ethical decision making? An empirical test of integrative social contracts theory, *Academy of Management Journal*, 47: 610-620.

Spini, D. 2003. Measurement equivalence of 10 value types from the

Schwartz value survey across 21 countries. *Journal of Cross-Cultural Psychology*, 34(1): 3-23.

Steenkamp, J. E. M., & Baumgartner, H. 1998. Assessing measurment invariance in cross-national consumer research, *Journal of Consumer Research*, 25: 78-90.

Stevens, S. S. 1951. Mathematics, measurement, and psychophysics, In S. S. Stevens (Ed.), *Handbook of experimental psychology*: 1-49, NY: John Wiley and Sons.

Stevens, S. S. 1968. *Measurement, statistics, and the schemapiric view*, Science, 161: 849-856.

Tay, C., Ang, S., & Van Dyne, L. 2006. Personality, biographical characteristics, and job interview success: A longitudinal study of the mediating effects of interviewing self-efficacy and the moderating effects of internal locus of causality, *Jounal of Applied Psychology*, 91: 446-454.

Thissen, D. 1991. *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory* [Computer software], Chicago, IL: Scientific Software International.

Thissen, D. 2001. *IRTLRDIF v. 2.02b: Software for the computation of the statistics involved in item response theory likelihood-ratio test for differential item functioning* [Computer software], Chapel Hill, NC: LL Thurstone Psychometric Laboratory.

Thissen, D., Steinberg, L., & Wainer, H. 1988. Use of item response theory in the study of group differences in trace lines, In H. Wainer and H. I. Braun (Eds.), *Test validity*: 147-169, Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Thomson, G. H., & Ledermann, W. 1939. The influence of multivariate selection on the factorial analysis of ability, *British Journal of Psychology*, 29: 288-305.

Thurstone, L. L. 1947. *Multiple-factor analysis. A development and*

expansion of the vectors of mind, Chicago, IL: University of Chicago Press.

Touw-Otten, F., Meadows, K. 1996. Cross-cultural issues in outcome measurement. In A. Hutchinson, E. McColl, M. Christie, C. Rittleton, (Eds.), ***Outcome Measurement in Primary and Out-Patient Care***: 199-208. Newark, NJ: Harwood Academic Publishers.

Truxillo, D. M.,. Bauer, T. N., Campion, M. A., & Paronto, M. E. 2002. Selection. fairness information and applicant reactions: A longitudinal field study, ***Journal of Applied Psychology***, 87: 1020-1031.

Vandenberg, R. J. 2002. Toward a further understanding of and improvement in measurement invariance methods and procedures, ***Organizational Research Methods***, 5(2): 139-158.

Vandenberg, R. J., & Lance, C. E. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research, ***Organizational Research Methods***, 3(1): 4-70.

Vandenberg, R. J., & Self, R. M. 1993. Assessing newcomers′ changing commitment to the organization during the first 6months of work, ***Journal of Applied Psychology***, 78: 557-568.

Voronov, M., & Singer, J. 2002. The myth of individualism-collectivism: A critical review, ***The Journal of Social Psychology***, 142: 461-480.

Westaby, J. D., & Lowe, J. K. 2005. Risk-taking orientation and injury among youth workers: Examining the social influece of supervisors, coworkers, and parents, ***Journal of Applied Psychology***, 90: 1027-1035.

Woehr, D. J., Sheehan, M. K., & Bennett, W. Jr. 2005. Assessing measurement equivalence across rating sources: A Multitrait-Multirater approach, ***Journal of Applied Psychology***, 90: 592-600.

Yen, W. M. 1993. Scaling performance assessments: Strategies for managing local item dependence, ***Journal of Educational Measurement***, 30: 187-213.

The Meaning and Verification Methods of Measurement Equivalence/Invariance

Won-Woo Park*
Yoon Hee Yang**
Hyun Jung Lee***
Yongjun Choi****
Moon Joung Kim**

ABSTRACT

Measurement equivalence/invariance(ME/I) has largely been ignored by researchers who examined cross-group differences in social science (e.g., human resources management, organizational behavior, marketing, and etc.). Thus, this study basically aims at investigating the meaning, importance and verification methods of measurement equivalence/invariance.

ME/I refers to the status that measurements in different conditions (e.g., time, region, culture, group, and medium) have comparable attributes. The establishment of ME/I is a critical prerequisite to ensure the validity of study for cross-group comparisons. In order to assess the researchers' awareness of ME/I we reviewed papers published in Korean Management Review, Korean Journal of Management, and Korean Personnel Administration Journal in last ten yearsand those in Academy of Management Journal and Journal of Applied

* Professor, College of Business Administration, Seoul National University
** Doctoral student, College of Business Administration, Seoul National University
*** Master student, Survey Methodology Program, University of Michigan
**** Doctoral student, Carlson School of Management, University of Minnesota

Psychology in last five years. Results suggested that researchers of the Korean journals significantly lack in the recognition of ME/I compared to foreign researchers. Based on the findings, the authors recommend Korean researchers understanding the importance of ME/I and verifying ME/I to improve validity of their studies.

Three frequently used practices to assess ME/I are proposed: (1) confirmatory factor analysis, (2) item response theory, and (3) generalizability theory. The concept and empirical application of each method are elaborated. Each of these three methods has its own pros and cons, and it is encouraged to use all the three in a complementary way. When it is practically not possible to combine all, researchers should choose a right method that best fits their own purpose.

Key words: Measurement equivalence, Measurement invariance, Confirmatory factor analysis, Item response theory, Generalizability theory