韓國心理學會誌
Korean Journal of Psychology
1996. Vol. 15, No. 1, 1-25

표준화 심리검사의 품질 문제 : 13개 검사도구의 기술적 품질 검토*

이 순 묵**

성균관대학교 산업심리학과

심리검사의 품질문제를 풀어가는데 있어서 표준화된 정도를 중심으로 접근하고자 한다. 표준화의 의미는 단순히 도구의 표준화가 아니라 제작, 실시, 채점, 해석, 및 의사결정에 사용하기까지의 전과정에서 표준화된 관행이 있을 때를 표준화검사(standardized testing)로 부르기로 한다. 이들 전과정에 대한 표준이 국내에서는 아직 구비되어 있지 않으므로 미국의 표준서(1985년 발행)를 참고하여 국내의 검사도구들을 기술적 품질의 측면에서 검토하고자 하였다. 검사의 기술적 품질에서 가장 많이 언급되는 부분이 타당도와 신뢰도라고 할 때, 신뢰도에 대한 개념은 비교적 보편화되어 있으나 타당도에 대해서는 그렇지 못하다. 검사에 종사하는 연구자 및 응용·사용자로서는 보다 적극적 접근을 통해서 지속적으로 타당화 과정을 추구해야 할 것이다.

심리학계에서 널리 쓰이는 Wolman(1989)의 행동과학사전에 의하면 표준화 검사는 아래와 같이 정의된다.

> 경험적으로 작성되어 실시 및 사용을 위한 분명한 지침이 있고 적절한 규준 및 신뢰도와 타당도의 자료가 제시되는 검사(test) (p.342)

따라서 표준화 정도를 판단하기 위해서는 실시 및 사용을 위한 지침에 포함되어야 할 내용 및 상세함의 수준, 규준, 및 신뢰도와 타당도 자료의 내용 및 수준에 대한 규정(표준)이 필요한데 우리나라의 심리검사분야에 보편화되고 사회적 구속력을 지니는 검사의 표준이 아직은 구비되어 있지 않다.

이종승(1987)은 한국심리검사 저작가 협회의 윤리위원회에서 1968년에 "심리검사기준"을 만들어 다음의 사항들을 심사하였다는 보고를 하고 있다 - 타당도, 신뢰도, 규준, 검사지 및 검사요강, 인쇄 및 체제 등. 그러나 그 위원회가 법적 집행력을 가진 형편이 아니었으므로 국내에서 발행되는 모든 심리검사에 대해 검사를

* 이 글은 1996년 2월 28일 한국 교육평가연구회의 심포지움에서 발표한 원고를 기초로 한 것임.
** 이 글의 초고에 대하여 많은 조언을 주신 두 분의 심사위원께 감사드린다.

실시하지도 못했으며 위원회 자체도 얼마 안가서 유명무실하게 되어 버렸다고 한다.

한편 한국심리학회에서는 1992년 10월에 심리검사심의위원회가 설치되면서 심리검사의 개발 및 사용에 대한 가이드 및 사회공익을 위한 검사관련활동을 목표로 출발하였으나 96년 10월말까지 그 임기가 4번 반복되는 동안에 이 위원회 역시 이렇다 할 기여를 하지 못하였다.

미국의 경우 이미 1954년에 심리학회에서 출간된 "심리검사 및 진단기법에 대한 기술적인 지침" 이래로 1985년에 개정된 지침이 (Standards for Educational and Psychological Testing) 교육학회, 심리학회, 및 교육측정학회의 지원아래 작성되고 심리학회에 의해서 출판되었다. 이것은 1954년 이래로 다섯번째의 지침이 되며 그 이전의 지침들을 대체한다.

이 글에서는 이 표준서를 잠정적인 준거로 삼아 심리검사의 품질이라고 하는 주제를 풀어나가고자 한다. 앞으로 이 표준서를 "검사 표준서" 또는 그냥 "표준서"로 언급하기로 한다. 표준서는 95년에 우리말로 번역(이순묵·이봉건, 1995)되었다.

표준서에서의 표준은 크게 나누어서 세가지로 나뉜다.

㉠검사제작 및 평가를 위한 기술적 표준
㉡전문분야 및 특별한 경우에 사용(활용)을 위한 표준
㉢실시절차에 대한 표준

이들 표준은 검사의 기술적 적합성, 사용상의 적절성, 그리고 추론상의 합리성을 보장하기 위한 것이다. 이 세가지 표준을 줄여서 기술표준, 사용표준, 및 절차표준으로 부르기로 한다. 표준화 검사라고 하면 이 세가지 표준이 충족된 검사를 의미한다. 즉 검사가 제작상의 기술적 표준을 충족시키는 것은 물론 사용분야에서 사용자가 지켜야 할 표준 및 검사실시 절차가 표준대로 실시될 때 검사가 목적한 바의 기여를 할 수가 있다. 즉 검사결과에 기초하여 올바른 추론을 도출할 수가 있다.

세 종류의 표준가운데 기술표준은 대체로 검사의 교본에 제시된 것을 검토하므로서 어느 정도 충족되는지 판단할 수가 있다. 기술표준의 내용은 타당도, 신뢰도와 측정오차, 검사개발 및 수정, 척도설정·규준설정·점수의 비교가능성·검사의 동등화, 끝으로 검사의 출판에 대한 것이다. 이들 내용은 검사교본에서 부분적으로 또는 상당한 정도로 제시되며, 제시된 내용을 보면 이들 표준이 그 검사에서 어떻게 얼마나 충족되고 있는지 판단할 수가 있다. 그러나 사용표준과 절차표준은 형편이 다르다.

우선 사용표준의 몇 가지 예를 보기로 한다. 번호는 표준에서의 번호이고, 번호 다음의 괄호안 내용은 저자가 요약해서 임의로 붙인 제목이다.

6.1 (사용시 자료평가)
검사사용자들은 특정한 목적을 위해 사용하려는 검사의 타당도와 신뢰도에 관한 인쇄된 자료를 가능한 한 모두 평가해야 한다.

6.3 (타당화되지 않은 목적)
어떤 검사가, 이전에 타당화 작업이 이루어지지 않은 목적을 위해 또는 타당도가 뒷받침 되지 못하는 용도를 위해 사용될 때에는 검사사용자가 타당도에 대한 증거를 마련할 책임이 있다.

위의 표준들을 볼 때 과연 어느 검사의 사용자가 표준 6.1과 6.3을 지키고 있는 지는 사용자의 보고서나 사용자와의 면접을 통하지 않고는 알 수가 없다.

절차표준은 그 충족되는 정도를 파악하기가 사용표준에 비해 훨씬 어렵다. 이 표준은 검사의 실시, 채점, 및 보고서 작성의 절차와 응답자의 권리보호를 위한 절차의 표준이다. 우선 표준서에서 몇 가지 예를 보기로 한다.

15.1 (표준적 절차 준수)
통상적인 경우에 있어서, 검사실시자는 검사출판자가 세부적으로 밝혀 놓은 실시 및 채점에 대한 표준화된 절차를 주의깊게 따라야 한다. 응답자에게 주어지는 지시사항, 시간제한, 문항제시방식, 그리고 검사자료나 장비에 대한 세부사항을 엄격하게 지켜야 한다. 이에 대한 예외는 전문적 판단을 조심스럽

게 내렸을 때에만 인정되며, 주로 임상적 적용시이다.

15.2 (검사환경)

검사환경은 아주 편안해야 하며, 주의를 산만하게 하는 정도가 최소한도여야 한다.

이상의 예에서 보듯이 절차표준은 사용표준에 비해서 훨씬 더 행동적이고 현장적이다. 즉 보고서를 통해서 그 충족여부를 판단하기보다 직접 실시현장에 가서 실시자의 행동을 관찰하므로서만이 충족여부를 판단할 수가 있다.

사용표준과 절차표준의 충족여부를 간단히 파악하기가 어렵다고 해서 그 표준들의 준수의무를 면제하고자 함은 아니다. 검사는 개인피검자로서는 일상생활에서 자신의 의견을 계량적으로 표현하는 매개체가 되기도 하고 인생의 주요 사건에서 갈림길을 좌우하는 의사결정의 준거가 되기도 하며, 올바르게 사용될 경우 사회적 공익의 도구가 되므로 그 제작에서부터 검사결과에 기초한 올바른 추론까지가 검사의 표준화에 대한 범위로 정의되어야 한다. 즉 여기서 저자는 검사의 제작에서 검사점수에 대한 추론까지의 전체과정이 기술표준, 사용표준, 그리고 절차표준에 비추어 심각한 결함이 없을 때 표준화된 검사(testing)라고 정의한다. 이것은 표준서에서 타당도의 개념이 "검사점수에서 도출된 특정의 추론이 적절한지, 의미있는지, 그리고 유용한 지에 대한 개념"으로 정의되는 것과 맥을 같이 한다.

여기서 저자는 표준화검사의 의미를 Wolman의 사전적 정의인 "standardized test"를 벗어나 "standardized testing"으로 확장하고 있으므로 잠시 "test" 또는 "검사"의 의미를 짚어보기로 한다. "test"라고 하면 영어에서 네 가지 의미로 볼 수 있다: logic of a test, test instrument, test score, testing. 이들 각각을 우리말로 하자면 검사논리, 검사도구, 검사점수, 및 검사관행이라고 할 수 있을 것이다.

검사논리: 지능검사, 적성검사, 또는 학력검사라고 할 때는 행동의 표본을 측정하는 논리나 절차라고 하는 추상명사로서의 검사를 의미하는 경우가 있다. 지능검사에 관한 토론회, 적성검사의 의미, 또는 학력검사의 기능 등이라는 말에서 우리는 지능, 적성, 또는 학력을 재는데 있어서의 또는 그 결과를 사용하는데 있어서의 논리나 절차를 의미한다.

검사도구: 어떤 논리나 절차에 따라서 손에 쥘수 있는 구체적인 도구가 만들어지면, 고대-비네검사, 노동부 추천의 직업적성검사, 어느 학년도 고등학생의 학기말 학력검사 등에서의 검사는 검사도구를 의미한다.

검사점수: 한편 검사의 신뢰도, 검사의 타당도, 또는 검사간의 상관 등과 같은 용어에서의 검사는 검사점수를 의미한다.

검사관행: 여기에는 좁은 의미와 넓은 의미가 있다. 검사담당자, 검사시간표, 검사일정, 또는 검사실시 등의 용어에서 검사는 좁은 의미에서의 검사관행을 의미한다. 검사관행을 좁게 보면 「검사실시」로 볼 수 있으나, 넓게 보면 검사의 논리설정, 그에 따른 검사도구제작, 검사의 실시, 검사점수계산, 검사점수의 해석 및 의사결정을 위한 사용까지의 전체과정을 검사관행(testing)으로 볼 수 있다. 검사표준서에서의 표준들 역시 이 전체과정에 대해서 취급하고 있다. 따라서 이 글에서도 좁은 의미의 검사관행은 '검사실시'로 부르고 '검사관행'이란 용어는 넓은 의미로 사용하기로 한다.

표준화검사가 단순히 잘 만들어진 검사도구를 의미하는 때도 있었다. 그 때는 검사의 타당도가 '재고자 하는 것을 재는 정도'로 정의되었었다. 그러나 요즈음 검사타당도의 정의는 단순히, 재고자 하는 것을 재는 정도를 의미하는 좁은 의미에서, 이제는 검사점수에서 도출된 추론의 적절성, 의미성, 및 유용성을 의미하는 넓고 동태적인 의미로 확장되어가고 있다. 즉 전에는 검사도구를 잘 제작만 하면 되었지만 이제는 잘 만들고 잘 써야 그 검사가 타당도있는 검사가 된다. 즉 잘 만들어진 검사도구가 "잘 사용되는 경우"를 거쳐야 타당도가 높아진다. 좋은 논리에 따라 잘 만들어지고 잘

사용되는 전체과정이 testing이다. 따라서 "표준화검사"의 의미도 이제는 Wolman(1989)의 정의인 standardized test instrument가 아니라 standardized testing으로 이해해야 할 것이다. 표준화된 검사관행이 없다면 표준화된 도구가 무슨 의미가 있겠는가? 즉 표준화검사라면 도구(test instrument)의 기술적 품질을 보장하는 규준, 타당도, 신뢰도 등이 만족스러운 것은(이것은 단지 필요조건일 뿐임) 물론 사용자를 위한 활용상의 지침과 실제에서의 실시·채점·보고작성·응답자 권리보호 등의 절차적 지침이 준수되는 검사전반의 관행으로 정의되어야 할 것이다. 이 글에서도 '검사'라는 용어는 맥락에 따라 논리, 도구, 점수, 또는 관행(넓은 의미)으로 이해될 것이다.

표준화검사에 대한 세 종류의 표준 중에서 기술표준이 특히 검사도구의 제작에 크게 관여된다. 전통적으로는 기술표준이 충족된 도구를 우리는 품질이 또는 양호도가 높은 검사도구라고 불러왔다. 그러나 어떤 도구가 기술표준을 충족한다고 해서 그 도구를 사용한 검사실시나 점수에 근거한 의사결정이 표준화되었음을 의미하진 않고 다만 필요조건이 충족되었다고 할 수 있을 것이다. 이 글의 목적은 국내에서 통용되는 검사도구들이 바로 이 필요조건을 얼마나 충족시키는 가를 보는 것에 중점을 둔다. 사용표준과 절차표준의 충족 정도는 경험적 조사의 여건이 허락치 않아서 여기서 논의할 형편이 안된다.

저자가 품질이란 용어를 쓰는 것은 이미 심리검사의 활용이 정착되어 있는 영미권의 학술지에서 심리검사의 질적 수준을 논의할 때 사용하는 "quality"라고 하는 용어를 번역한 것이다. 품질이란 결코 물리적 제품에 대한 것만 아니라 서비스(service)에 대한 품질도 논의될 정도로 그 의미가 확장되고 있으므로 심리검사의 "quality"를 "품질"로 번역함에 무리가 없다고 보았다. 심리검사의 품질에 대한 논의가 국내에서 명시적으로 제기되기는 교육학 분야에서이다(예: 이종승, 1987). 교육학 분야에서는 「품질」이란 용어 대신에 「양호도」란 용어를 사용해 왔다. 교육학 문헌에서 심리검사의 양호도라고 하면 대체로 네 개의 개념, 즉 타당도, 신뢰도, 객관도, 및 실용도를 의미한다. 앞으로 이 글에서 품질 또는 양호도라고 하면 심리검사의 품질 또는 양호도를 의미하기로 한다. 품질 좋은 검사가 가져야 할 특성으로서 타당도와 신뢰도를 각각 하나씩의 주제로 언급하는 점에서는 어떤 문헌에서도 일관성이 있다. 그러나 객관도와 실용도에 대해서는 문헌에 따라 상이한 입장을 보인다. 이종성·강봉규 및 한종철(1982)은 이 두 가지를 모두 명시적으로 언급하고 있다. 검사표준서에서는 이 두 가지를 명시적으로 언급하지는 않고 있다. 단지 신뢰도의 章에 있는 표준 2.8에서만 객관도를 언급하고 있을 뿐이다. Thorndike, Cunningham, Thorndike 및 Hagen(1991)은 검사의 품질을 평가하는데 세 가지를 고려한다면 그것은 신뢰도, 타당도, 및 실용도라고 명시하고 있다(p.91, 146-151 참조). 그러나 객관도에 대해서는 평정척도의 경우에 적용되는 평정자간 신뢰도라는 소제목으로 다루고 있다.

품질의 네 가지 개념을 검사표준서와 연결해 보면 타당도, 신뢰도, 객관도는 기술표준의 주된 내용이 된다. 실용도에 대해서는 특별히 꼬집어서 언급은 안하지만 기술표준, 사용표준 및 절차표준이라고 하는 전체를 통틀어서 달성하려고 하면 충족되게 된다. 실용도에서 중요한 내용을 보면 경제성, 검사실시의 편리성, 해석 및 점수사용의 편리성(Thorndike 등, 1991)이다. 경제성은 금전적 경제, 시간의 경제 또는 채점의 용이함인데 이것은 검사가 기술표준, 사용표준 및 절차표준에 맞게 작성되고 활용될 때 고객에게는 품질에 비추어 경제성있는 검사로 인식될 수가 있는 것으로 이해될 수 있다. 검사실시의 편리성은 지시사항이 분명하고 완전하면 절차표준이 쉽게 준수되므로 어느 정도 확보된다. 해석 및 점수사용의 편리성 역시 검사가 무슨 기능을 위해서 개발되었는가와 어떻게 개발되었는가의 일반적 절차가 잘 기술되

면(검사출판에 관련된 표준) 상당한 정도 확보된다. 따라서 검사의 품질은 넓게 볼 때 검사의 표준이 지켜지는 정도이다.

그렇다면 검사의 품질은 검사관행(넓은 의미)의 품질이며 세 가지로 크게 보면 기술적 품질, 사용상 품질, 그리고 절차상 품질을 논의할 수 있다. 기술적 품질은 곧 검사도구의 품질을 의미하며 보다 세분하면 검사논리의 품질, 검사외관의 품질(안면타당도), 검사문항의 영역(domain) 대표적인 품질(내용타당도), 검사점수의 품질(신뢰도, 구성개념타당도, 준거타당도, 측정오차, 문항이나 검사가 제공하는 정보량), 검사점수에 대한 의미부여의 품질(척도설정, 규준설정, 점수의 비교가능성, 검사동등화), 그리고 교본(manual)의 품질(완전성, 정확성, 명확성)로 나누어 볼 수 있다. 그런데 검사도구의 품질이 자동차나 비행기 등 물리적 제품의 품질과 다른 점은 품질을 평가할 수 있는 현실이 전자는 사회적 현실이고 후자는 물리적 현실이라는 것이다. 물리적 현실에는 물리적인 기준이 있어서 자동차의 경우 BMW라고 하면 언제 어디에 내놔도 품질좋은 차라는 이야기를 듣는다. 그러나 사회적 현실은 똑같은 제품이라해도 그 상황과 적용되는 대상에 따라서 품질이 상대적으로 결정되는 것이다. 심리검사도구가 바로 그렇다. 어느 심리검사도구를 언제 어디에 적용해도 높은 타당도를 가진다고 이야기하지 않는다. 타당도가 높은 상황과 적용대상이 제한되므로 기술적 품질 역시 상대적이다. 따라서 자동차의 품질과 심리검사도구의 품질은 다른 개념이다. 전자는 사용이 어디에 되느냐를(예: 고급승용차, 스포츠카로, 또는 용달차로) 불문하고 품질을 논할 수 있으나 후자는 '어떠한' 사용이 전제되지 않고는 품질이 논의될 수가 없다. 그래서 검사도구의 품질은 검사의 품질 즉 검사관행의 품질을 구성하는 한 부분이 될 뿐이고 검사관행의 품질과 별도로 독립적인 개념이 될 수 없다는 것이 저자의 견해이다.

이제 이 글에서 국내의 검사도구들에서 실제로 이들 표준이, 특히 기술표준이 얼마나 준수되고 있는가를 살펴보므로서 우리가 알고 있는 '표준화' 검사들의 품질의 일면, 즉 기술적 품질을 판단해 보고자 한다. 기술적 품질의 판단은 검사의 교본에 대한 검토를 우선적으로 요구한다. 그리고 그 검사가 제작된 이후 추후연구되어 보고된 내용들 역시 검토에 포함되어야 한다. 물론 검사교본 제작시에 제작자들은 그 검사를 사용하여 연구된 결과에 대한 내용들을 포함해야 한다. 그러므로서 사용자로 하여금 그 검사의 품질수준을 판단할 수 있게 해야 한다. 그러나 우리의 현실에서는 시간적, 경제적 제약 때문에 일단 제작부터 하고 그 이후에 그 검사를 사용한 연구가 되는 경우가 더 많다고 해야할 것이다. 그렇다면 교본의 제작자들은 검사제작후 적어도 5-6년 후에는 검사의 품질에 대한 정보를 제공할 수 있는 추후연구 및 결과를 교본에 소개해서 새로운 교본을 사용자들에게 제공해야 할 것이다. 이 글에서 검토된 검사들의 경우 몇 개는 검사제작후(규준제작기준) 4-7년 된 것도 있고 나머지는 10년 이상된 것들이다. 이제는 그 동안의 추후연구에 기초한 새로운 교본제공의 시점이거나 그 시점을 지났다고 할 수 있다. 이 글에서는 일단은 저자가 접근가능한 교본을 중심으로 각 검사들의 품질을 판단하고 있으므로 추후연구들까지 반영하는 품질검토는 아님을 분명히 한다.

검사의 검토

저자가 검토할 수 있는 검사의 수에는 제한이 많았다. 저자는 최근에 많이 쓰이고 있는 검사를 수집하기 위해서 서울의 어느 사립대학과 지방 어느 국립대학의 학생생활연구소에서 빈번하게 사용되는 검사들 가운데 중복되지 않는 것들을 합하여 13가지를 검토하게 되었다. 이 중에서 몇 가지는 이미 이종승(1987)의 연

표1. 지능검사의 검토

기재사항	K-WAIS	KEDI-WISC	일반지능검사	지능성숙검사
제작자	염태호 등	박경숙 등	정범모, 김호권	김해옥 등
발행자	한국가이던스	도서출판 특수교육	코리안테스팅센타	교육과학사
규준제작년도 (자료수집기간)	1992 (91.8-92.6)	1991 (91.6-91.7)	1967 (65-66)	1963 (62.5-63.2)
규준제작후 경과	4년	5년	29년	33년
규준	7개 연령집단	11개집단 (5세~15세까지)	2개 (검사유형 A, B에 대한 대학생집단)	16개 (지역별, 연령별)
표집방식	인구조사자료(1989)를 이용 연령, 성별, 지역 거주지, 학력 고려한 층화표집	성, 인구, 사회경제적 연령, 지위 등을 고려한 층화표집	전국대학에서 층화표집	X
총문항수	168개	155개	110개	94개
규준집단의 크기	200명내외	190~221명	A형검사: 1634명 B형검사: 1597명	각 규준집단별로 보고는 안했으나 지역별 보고한 것을 기초로 추정하면 500명이상은 될것임
관찰-문항비율 (규준집단크기 ÷총문항수)	1.2배	최고 1.4배	A형: 14.9배 B형: 14.5배	대략 5배 이상
신뢰도 보고구분	연령집단별 소검사별	연령집단별 소검사별	소검사별(대학생뿐이므로 연령구분 불요)	연령집단별(지역별은 X, 소검사별은 X)
신뢰도 계산방식	대체로 반분법. 속도검사나 두 개의 독립된 부분으로 될 경우 재검사법 (1-7개월 간격)	반분법. 재검사법(2주 간격)	X	KR-20 반분법
측정의 표준오차 보고	0	0	X	0
신뢰도값	최소인 것이 .70	반분신뢰도: .5이하인 것도 있음. 재검사법: .12인 것도 있음.	.68~.88	.90 이상
타당도증거 1. 내용타당도 2. 준거타당도 3. 구성개념타당도	WAIS X 소검사간 상관	WISC 고대-비네검사와 상관 (값이 .04, .16, .17, .21등과 같이 작은 것도 있음) 소검사간 상관	X X 소검사간 상관	X 연령을 준거로 연령과 검사점수와 상관 X
검사실시·해석하는 사람의 자격	전문적인 소양과 경험 갖출 것을 언급. 구체적 명시가 필요.		X	X
검사의 과학적 기초에 대한 증거수집	Wechsler의 이론모형 사용. 그러나 지난 4-5년간 국내에서의 후속연구에 대한 보고가 추가되어야 할 것임.		X	X
논리의 철저한 명 및 용도를 지지하는 자료와 연구	검사논리에 대한 설명충분. 자료와 연구용도에 대한 추천을 지지하는 자료없음		X	X

구에서 검토된 것이지만 이 글에서는 좀 더 세부적으로 살펴가며 그 품질을 검토해 보기로 한다. 이 글에서 검토되는 13가지의 검사는 지능검사 4개, 성격검사 4개, 적성검사 3개, 임상검사 2개이다. 그러면 각 검사 종류별로 검토를 제시하기로 한다.

각 검사의 검토에 들어가기 전에 이 글에서는 검사받는 사람들 또는 검사도구에 응하여 자신에 대한 정보를 제공하는 사람들을 "응답자"라고 언급하기로 한다. 검사도구라고 하면 표준서에서는 구조화된 과제, 설문, 그리고 행동표본으로 나눈다(이순묵·이봉건, 1995, p.21-23). 각 도구에 응하므로서 검사점수의 원천이 되는 사람들을 검사실시의 맥락에 따라서 다르게 부른다. 즉, 구조화된 과제일 경우 응시자(시험일 경우), 내담자(임상, 상담장면), 또는 피검자, 수검자 등으로 부르게 되지만 설문일 경우 단순히 응답자라고 하는 것 이외에 내담자, 피검자, 수검자 등 어느 것도 적당하지 않다. 또한 행동표본일 경우 응답자일 수도 있고 피관찰자일 수도 있지만 내담자, 피검자, 수검자 라는 용어는 크게 해당되지 않을 것이다. 따라서 검사에 응하여 또는 그 경우에 응하여 검사점수의 원천이 되는 사람들을 이 글에서는 통틀어 "응답자"로 부르기로 하고 맥락에 특수한 용어는 제한적으로 꼭 필요한 경우에만 사용하기로 한다.

지능검사

우선 세부적인 검토를 하고 검토에 대한 요약을 하기로 한다.

(1) 세부검토

표 1과 같이 4개의 검사를 검토하였다: K-WAIS(염태호, 박영숙, 오경자, 김정규, 이영호, 1992), KEDI-WISC(박경숙, 윤점룡, 박효정, 박혜정, 권기욱, 1991), 일반지능검사(정범모, 김호권, 1993), 지능성숙검사(김해옥, 김남수, 임의도, 이인수, 조석호, 1993)

여기서 괄호안의 이름은 저작권자보다는 실제 제작자들의 이름을 기재한 것이며 연도는 저자가 가지고 있는 교본 또는 실시요강의 발행일자이다. 검사에 따라서는 제작자와 저작권자가 다르지만 제작자들의 기여를 인지하는 의미에서 제작자를 밝히고자 한다. 또한 어떤 기재사항에 대해서 O는 보고하였음을, X는 보고하지 않았음을 의미한다. 이와 같은 방식은 이 글에서 다른 종류의 검사를 검토함에 똑같이 적용된다.

(2) 지능검사 검토 요약

ㄱ. 규준제작을 위한 자료수집기간이 보통 1년씩 걸리고 있다. 좀 더 짧은 시간에 자료가 수집되는 것이 자료수집에 있어서 오염변수의 개입을 막을 수 있을 것이다.

ㄴ. 규준제작후 경과기간

이 기간을 보면 일반지능검사(정범모, 김호권, 1993)와 지능성숙검사(김해옥등, 1993)는 30여년씩이나 된다. 그 동안에 시대가 변하여 어린이들이 보고 듣는 경험의 변화가 대단하다. 따라서 척도체계가 30년전과 지금은 변했을 것으로 본다. 시간이 흐름에 따라 동일한 척도체계를 유지하는 것이 바람직하다. 척도체계의 안정성(stability of scale)에 대한 정기적인 검토를 해야 한다(표준 4.9). 척도체계는 원점수(raw score)에 비교정보를 포함시키기 위해서 전환시킨 점수, 즉 척도점수(scale score)로서 주어진 값들이 바로 척도체계이다. 많은 지능검사의 경우 원점수에 대한 척도점수로서 T점수를 사용한다. 척도체계의 안정성은 검사제작시에 어떤 원점수에 어떤 척도점수가 주어졌다면 시간이 지난 후 현재도 같은 원점수에 같은 척도점수가 부여되는 것이 타당한 정도를 의미한다. 예로서 지능검사를 보자. 지능수준은 자기또래 아이들을 규준집단으로 할 때 평균의 위치에 있는 아이가 척도상으로는 지능지수 100을 부여받게 된다. 30년전에 X개를 맞은 아이가 지능지수 100을 받았으면 같은 도구를 사용해서 오늘날 아이들에게 사용할 경우에도 X개 맞은 아이에게 지능지수 100을 주는 것이 타당하다면 척도체계는 안정되어 있는

것이다. 그러나 적어도 (x+a)개를 맞은 아이에게 지능지수 100을 주는 것이 타당하다면 척도체계는 이미 변한 것이다. 원점수(raw score)로 볼 때 응답자들의 평균과 표준편차 중 하나 또는 둘 다 변하면 척도체계는 경험적으로 이미 변한 것이다. 이 글에 검토된 두 지능검사는 이런 점에서 척도체계의 안정성을 검토하고 아직도 규준과 척도체계가 유효한 지에 대한 최신 정보를 사용자들에게 공급해야(표준 5.5) 할 것이다. 즉 규준의 최신성(이종승, 1987에서는 근대성으로 부르고 있음)을 유지해야 할 필요가 있다.

ㄷ. 규준은 실제로 검사를 실시해서 나온 점수를 통상적으로 비교하고 싶어하는 표적이 되는 집단에서의 점수분포이다. 규준은 사용자들이 검사를 사용하는 여러가지 상황을 고려하여 여러 종류가 있는 것이 바람직하다. 즉 규준은 사용자의 상황에 따라 상대적인 것이다. 따라서 검사개발자들이 충분히 세분된 규준을 제공하지 못하는 경우 사용자들이 자신의 상황에서 실제 사용에 있어 적합성(이종승, 1987)이 저하될 수 있다. 그럴 경우 개발자로서는 사용자들이 국지적 규준을 개발할 것을 권장해야 한다(표준 4.3). 일반지능검사에서는 2개밖에 규준이 없는 것을 보완하기 위하여 사용자들에게 자기 학교의 규준을 작성할 것을 권장하고 있는데 이것은 올바른 권고이다. KEDI-WISC의 경우 지역별(예: 도시, 지방), 또는 "남아/여아"별 규준이 없는데 혹시 이러한 규준을 사용자가 필요로 할 경우는 없을지 궁금하다.

ㄹ. 규준작성을 위한 표본추출의 설계, 참여도

규준작성시 표본추출의 과정과 표본자체에 대한 명시적 제시가 있어야 규준집단의 대표성과 적합성을 판단할 수 있다. 이러한 정보를 상술해야(표준 4.4) 연구의 적절성을 평가할 수 있는데 대체로 층화표집을 했다고 전반적인 보고를 하고 있으나 각 층별 구분에 대한 보다 자세한 보고가 아쉽다. 지능성숙검사의 경우 전혀 표집방식에 대한 보고가 없어서 규준의 대표성과 적합성(이종승, 1987)을 확인할 수가 없다.

ㅁ. 관찰-문항비율

이것은 규준개발을 위한 통계처리를 위해서 최소한으로 필요한 표본을 확보할 수 있는 표본설계(표준 4.4)인지를 판단할 수 있게 해준다. 규준개발시에 사용하는 전형적인 통계처리는 상관분석, 회귀분석, 요인분석 등이다. 이 중에서 요인분석은 척도의 1차원성, 수렴타당도, 그리고 변별타당도 등을 보이는데 거의 필수적인 통계처리이다. 요인분석시에 변수에 대한 관찰의 배수는 바람직하게는 10배(Nunnally, 1978) 최소한도 5배(Gorsuch, 1974)는 되어야 한다. 물론 자료내의 구조가 특히 분명할 경우 5배보다 작을 수도 있지만 표본이 클수록 자료분석의 결과에 안정성이 있으므로 아무리 작은 규준집단이라 해도 200명은 넘어야 요인분석을 함에 무리가 없을 것이라는 것이 Guilford(1954)의 오랜 경험에서 나온 주장이다. 이러한 여러 가지 견해를 종합해 볼 때 K-WAIS[1]와 KEDI-WISC는 표본의 크기가 너무 작지 않나 생각된다.

ㅂ. 신뢰도와 측정오차의 보고구분

부분점수, 구성점수(composite score), 총점수 등에 대해 모두 신뢰도와 측정의 표준오차를 정확히 보고해야 한다(표준 2.1). 이것은 소검사별로도 신뢰도를 보고할 것을 의미한다. 소검사끼리는 측정되는 구성개념도 다르므로 신뢰도가 같을 이유가 없기 때문이다. 또한 신뢰도나 측정의 표준오차가 상이한 규준집단간에 현저하게 다를 것으로 기대되는 이유가 있으면 각각의 주요 규준집단에 대해서 따로따로 값을 보고해야 한다(표준 2.9). 이런 관점에서 볼 때 K-WAIS와 KEDI-WISC는 아주 모범적으로 보고를 하고 있다. 일반지능검사에서는 대학생 전체가 규준의 대상이므로 소검사별로만 보고하고 있는 것은 올바른 방식이지만 측

1). K-WAIS에서 "바꿔쓰기" 부분은 93개의 문자가 있었지만, 이 글에서 문항수효를 계산할 때는 1개의 문항으로 취급하여 '관찰-문항비율'의 산출에 사용하였음.

정의 표준오차가 보고되지 않은 것은 유감이다. 지능성숙검사에서 소검사별로 신뢰도를 보고하지 않는 것은 지적되어야 할 사항이다. 지능성숙검사에서 또한 지역별 규준집단별로 신뢰도가 모두 같다는 가정하에 지역별 신뢰도는 없는데 과연 그럴지에 대한 논의가 제시되면 더욱 신빙성있는 보고가 될 것이다.

ㅅ. 신뢰도 계산방식

신뢰도 계산방식은 분명하게 정의되어야 한다(표준 2.3). 지능검사는 대체로 시간제한이 있는 속도검사이다. 속도검사일 경우 신뢰도의 값이 과대 계산될 수 있는 방식으로 신뢰도를 계산해서는 안된다(표준 2.7). 즉 내적 일관성법(예: 반분신뢰도, α계수, KR-20)으로 계산하면 과대 계산될 수가 있다. 일반지능검사의 경우 아예 계산방식이 제시되지 않았지만 나머지 세개의 검사에서도 내적 일관성의 신뢰도 계수를 보이고 있다. 물론 K-WAIS와 KEDI-WISC에서는 재검사법도 사용하고 있으나 좀더 전반적으로 재검사법을 사용하든가, 검사를 평행한 두 부분(1부, 2부)으로 나누어 1부를 먼저 시행한 후 2부를 시행하여 그 두 부분검사의 점수간 상관을 구하여 Spearman-Brown 공식으로 수정하여 신뢰도를 구하는 것도 한 방법일 것이다.

ㅇ. 신뢰도의 값

KEDI-WISC에서는 특별히 신뢰도 값이 낮은 경우가 있는데 그 만큼 검사의 품질을 잠식하는 면이 된다. 신뢰도의 값들은 높을수록 좋다. 공식적으로 판매되어 많은 사람의 인생을 좌우하는 판단에 사용될 '표준화검사'라면 더욱 그러하다. Mehrens와 Lehmann(1980)은 개인에 관한 의사결정에 사용될 경우 .85이상, 집단에 관한 의사결정에 사용될 경우 0.65이상의 신뢰도를 권하고 있다. 그러나 신뢰도값이 아무리 높게 제시되어 있어도 신뢰도를 구하는 집단의 표본크기 및 표집절차(표준 2.2)가 충실하지 않을 경우 그 값에 대한 신뢰가 떨어진다. 예로서 K-WAIS에서 3개의 연령집단의 각각에 대하여 30-40명을 표본으로 한 것은 결코 충분한 표본이라고 볼 수가 없다. 즉 그렇게 작은 표본에서 구해진 신뢰도값이 과연 같은 규모의 다른 표본에서도 일관성있게 나올 수 있는지에 의문을 가져볼 수 있다.

ㅈ. 내용타당도

검사의 내용이 과연 재고자 하는 개념의 영역을 잘 대변하고 있는지에 대해서는 거의 언급이 없다. Sternberg(1985)의 맥락적 지능이론(contextual theory)에서는 개인생활에 관련된 현실환경에 목적적으로 적응하고, 환경을 선택 또는 조성하는 활동을 지능으로 본다. 그 현실환경은 곧 문화차이를 함의한다. 따라서 같은 지능이라해도 문화에 따라 그 지능의 영역(domain)은 똑같을 수 없을 것이다. 그런 관점에서 볼 때 K-WAIS와 KEDI-WISC같이 외국의 것을 번안한 경우 내용타당도를 원본검사의 저명도에 어느 정도 의존할 수야 있겠으나 문화가 다른 우리나라에서 번안하는 과정에서 내용을 대표하기 위해 기울인 조심스런 노력에 대해 구체적인 언급을 해야 할 것이다. 지능성숙검사의 경우 규준표에는 7세, 8세에 대해서만 있는데 본문(p.15)에서는 6세부터 8세까지에 대한 검사라고 하고 있음은 수정되어야 할 것이다.

ㅊ. 준거타당도

외부준거로서 기존의 유사한 지능검사점수 또는 학교성적 등을 사용할 수가 있을텐데 KEDI-WISC와 지능성숙검사만이 약간의 노력을 보였을 뿐이다. 특히 지능성숙검사의 경우 연령에 따라 지능성숙이 진행한다는 이유에서 연령을 준거로 썼으나 이 경우는 불행히도 연령이 7세~8세 밖에 되지 않아서 타당도 계수의 크기는 '범위의 축소'에 의한 과소계산의 우려가 있을 것이다. 보다 충실한 논의를 거쳐 개발한 준거변수들에 대한 설명을 정확하게 제공하고(표준 1.12) 준거타당도를 보여야 할 것이다.

ㅋ. 구성개념타당도

구성개념타당도를 보이기 위한 노력은 소극적 노력과 적극적 노력으로 나누어 볼 수가 있

다. 전자는 그저 도움되는 정보는 아무 것이나 다 수집하자는 노력이고 후자는 미리 이론적으로 그 변수가 다른 변수들과 가지는 관계를 예측하여 자료에 비추어 그 관계를 검증하므로서 구성개념타당도를 확보하는 노력이다. 소검사간 상관을 보여서 유사한 검사들이 공변함을 보이고(표준 1.9, 수렴타당도) 관계가 먼 변수들이 공변하는 정도가 적음(표준 1.10, 변별타당도)을 보이는 것은 소극적 접근의 하나이다. 그런데 지능성숙검사에서는 그나마 이 방법도 취하지 않았고 설사 K-WAIS, KEDI-WISC, 일반지능검사에서 이 방법을 취하긴 했어도 그 상관을 보고 수렴성과 변별성에 대한 논의를 제공하지 않으므로서 실제 타당도의 증거를 보이는데 기여하지 못하였다. 앞으로 모든 검사에서 보다 적극적인 타당도 확보의 노력을 보여야 할 것이다.

ㅌ. 검사를 실시하는 사람의 자격

검사교본에서는 검사의 실시 및 적절한 해석을 위해 필요한, 특별한 자격을 훈련, 경험 및 자격증 등의 측면에서 구분해 놓아야 한다(표준 5.4). K-WAIS와 KEDI-WISC는 평가에 있어 평가자의 판단이 많이 사용되며 검사중 응답자의 행동이나 태도가 검사점수 해석을 위한 하나의 정보가 된다. 따라서 상당한 정도의 전문성(예: 관련된 대학원 석사과정 수료이상)이 필요하며 그러한 요건이 교본에 명시적·구체적으로 언급되어야 함에도 불구하고 교본에서는 피상적으로 '전문적인 소양과 경험'을 갖추어야 한다고만 언급하고 있다. 일반지능검사와 지능성숙검사는 자격요건에 대해 아예 언급이 없는데 지능검사점수가 피검자나 학부모들에게 미칠 수 있는 영향을 감안할 때, 최소한도의 자격요건에 대한 언급은 필요할 것이다.

ㅍ. 검사의 과학적 기초

검사 및 검사프로그램은 건전한 과학적 이론에 의거하여 개발되어야 하므로 개발자들은 그에 대한 증거를 수집해야 한다. 또한 무슨 정보가 필요한지를 파악하고 필요하다면 연구도 해야 한다(표준 3.1). 제작한 후 여러 연구를 거쳐 과학적 기초를 수집한 후 그 결과를 기초로 제작된 검사의 공식적 출판에 들어가는 것이 원칙이지만 그렇지 못한 우리나라의 검사제작문화에서는 검사제작에 이어 보완하는 후속연구의 필요성만이라도 촉구되어야 한다. 즉 자신이 개발한 검사의 과학적 기초를 보강하기 위한 연구가 필요하다. K-WAIS와 KEDI-WISC는 과학적 기초를 Wechsler의 지능이론모형에서 찾을 수 있겠으나 국내에서 지난 4-5년간에 누적된 연구결과를 교본개정시 소개하는 것이 바람직하다. 다른 두 지능검사에서는 이 점에 대한 노력이나 보고가 없다. 어느 지능검사의 요강을 보면 "이 요강은 임시로 내어 놓은 것으로 · · · 빠른 시일내에 상세한 · · · 보충된 요강이 나올 것을 약속하는 바이다" 하고서는 30년이 지나는 경우도 있는 것이 우리의 현실이니 제작자는 자신의 작품을 아끼는 마음으로 과학적 기초를 마련하는 노력을 기울여야 할 것이다.

ㅎ. 논리의 설명, 용도를 지지하는 자료 및 연구제공

교본에서는 검사의 논리에 대한 철저한 설명은 물론 용도에 대한 추천, 그 추천을 지지하는 자료의 요약이 제공되어야 한다(표준 5.2). 대체로 K-WAIS와 KEDI-WISC는 검사의 논리에 대해 충분한 해설을 하고 있다. 그러나 위의 네 가지 지능검사 모두가 그 교본에서 추천하는 용도에 대한 추천을 지지하는 자료의 제공은 전혀 없다. 국외 국내에 있을 법한 자료를 추가하는 것이 사용자들에게 보다 품질높은 검사임을 보여주는 한 방법이 될 것이다. 특히 일반적 사용 및 특수한 목적으로의 사용에 대한 국내 국외의 대표적인 연구들을 균형있게 언급해야 할 것이다(표준 5.3). 그런데 교본들에 그러한 자료 및 연구에 대한 제공을 하는 것을 거의 볼 수가 없으니 스스로 만든 제품의 품질을 보일 수 있는 기회를 방치하는 것이다.

성격검사

(1) 세부검토

여기서 검토된 4개의 성격검사는 다음과 같다: MBTI(김정택, 심혜숙, 1991), 성인용 성격요인검사(염태호, 김정규, 1990), 성격진단검사(이상로, 변창진, 진위교, 1979), 자아실현검사(김재은, 이광자, 1983). 이들 검사에 대한 검토 역시 지능검사에서 사용된 것과 같은 방식이다. 검토결과는 표 2를 사용해서 제시하기로 한다.

(2) 성격검사요약

ㄱ. 규준제작후 기간이 30년 가까이 되는 것도 두 개나 있다.

ㄴ. 규준의 수효가 적절한지에 대한 논의가 어느 검사에서도 없긴 하지만 검사의 목적별로 과연 사용자들에게 충분한 서비스가 될 정도의 수효가 갖추어졌는지 조사할 일이다.

ㄷ. 표집방식에 대해서 층화표집, 랜덤표집이라고 하지만 랜덤性과 층화의 성격을 나타내는 구체적인 표본설계를 밝히지는 않고 있음

ㄹ. 관찰-문항비율이 MBTI, 성격진단검사, 자아실현검사에서는 대단히 작다. 적어도 검사개발시에는 이 비율이 이렇게 작으면 안될 것이다. MBTI의 경우 규준집단의 크기가 41명밖에 안되는 경우도 있는데 이렇게 작은 크기는 관찰-문항비율을 나쁘게 할 수 밖에 없다. 또한 결코 규준으로서의 대표성이 있는 크기는 아니다. 그렇다면 규준의 적합성도 역시 의문의 대상이다.

ㅁ. 신뢰도 보고 구분

MBTI의 경우 E, I, · · · 등 8개의 개별척도에 대하여 점수를 매기고 있으나 그들에 대한 신뢰도를 보고하지 않았다. 성인용성격검사에서도 2차요인들이 구성점수(composite score)인데 그에 대한 신뢰도가 제시되지 않았다. 구성점수는 그 부분점수보다 신뢰도가 낮거나 높을 수가 있으므로 별도로 보고하여야 한다. 또한 각 인구집단간에 신뢰도가 현저하게 다르지 않을 때 한 가지로 신뢰도를 보고하는 것인데, 현저히 다른지 아닌지의 논의가 안되고 있음은 모든 검사에 공통적이다. 예로서 자아실현검사에서 규준집단은 대학생 남자/여자, 고등학생 남자/여자의 4집단인데 신뢰도는 대학생 100명에 대해서 고등학생 100명에 대해서만 구하였다. 만일에 대학생 남자집단과 고등학생 여자집단간에 신뢰도가 현저하게 다르다면 이 점을 간과한 것은 사용자에게 부정확한 정보를 주게 된다.

ㅂ. 신뢰도계산방식

재검사법을 쓴 경우 시간간격을 보고해야 하는데 MBTI와 성인용 성격검사에서는 그에 대한 보고가 없다. 특히 MBTI의 경우 응답자의 성격을 16개 유형에 분류해 넣는 것이므로 준거참조검사로서의 사용방식이 된다. 이 때 교본에서 제시하는 성격분류 방식이 의사결정의 일치도를 보장하는 정도가 검사의 신뢰도가 된다. MBTI에서는 전통적인 신뢰도는 물론 의사결정의 일치도를 함께 보고함이 바람직하다. 즉 개인이 또는 친한 친구가 자신에 대해서 생각하는 유형과, 실제 검사결과 나온 유형이 일치하는 정도를 보여야 할 것이다.

ㅅ. 측정의 표준오차

4개의 성격검사 중 어느 것도 이것을 보고하지 않고 있다.

ㅇ. 신뢰도의 값

Mehrens와 Lehmann(1980)의 기준(개인에 대한 의사결정에서 0.85이상)에서 보면 모두가 바람직한 수준은 아니다. 성인용성격검사의 경우 α계수가 매우 낮은 경우가 많은 대신 반분신뢰도와 재검사신뢰도는 높다. 검사가 반분될 때, 두 부분간 "평행"하다는 자신이 있으면 α계수보다는 반분신뢰도를 신뢰도값으로 보아야 할 것이다. 반분신뢰도계산시 이 "평행"의 가정을 어느 검사에서도 검토하지 않고 있음 역시 언급되어야 할 사항이다.

ㅈ. 내용타당도

성격진단검사를 제외하고는 모두 외국의 것을 번안한 것이므로 내용타당도를 크게 언급하

표2. 성격 검사의 검토

	MBTI	성인용성격검사	성격진단검사	자아실현검사
제작자	김정택, 심혜숙	염태호, 김정규	이상로, 변창진, 전위교	김재은, 이광자
발행자	한국심리검사연구소	한국심리적성연구소	중앙적성출판사	중앙적성
규준제작년도 (자료수집기간)	1990 (1987-1990)	1990 (?)	1979(?) (?)	1977 (?)
규준제작후 경과	6년	6년	27년(?)	29년
규준	15개	20개	6개	4개
표집방식	서울과 지방에서 고등학생, 대학생, 일반인을 랜덤표집	성별, 연령별, 학력별 분포 감안한 층화표집	지역, 학년, 학과, 남녀를 단위로 무선표집	전국적으로 무선표집
총문항수	94	165	350	130
규준집단의 크기	41명~1782명	70명~695명	987명~1268명	288~575명
관찰-문항비율	2배가 안되는 표본이 3개, 5배가 안되는 표본이 9개임(위의 3개 포함)	2배가 안되는 표본이 14개, 5배가 되는 표본은 없음	2배는 모두 넘고 5배가 되는 표본은 없다	2배는 모두 넘고 5배가 되는 표본은 없다
신뢰도 보고구분	① EI, SN, TF, JP에 대해서만 보고 ② 각 개별척도에 대해서는 보고 안함. ③ 인구집단별 신뢰도가 다를 수 있는 가능성의 논의없음	① 소검사별 보고 ② 규준집단간에 신뢰도가 다를 수 있는 가능성의 논의없음. ③ 1차요인을 결합해서 나오는 6개의 2차요인에 대한 신뢰도도 보고할 것(여기서 2차요인은 구성점수)	① 소검사별 보고 ② 규준집단별로 신뢰도가 다를 수 있는 가능성의 논의없음	① 소검사별 보고 ② 여자집단, 남자집단으로만 보고 ③ 대학생, 고등학생간에 신뢰도가 다를 수 있는 가능성의 논의 없음. ④ 신뢰도를 구할 때의 표본은 100명(작은 편임)
신뢰도 계산방식	① 재검사법 사용 (기간보고 X) ② 반분법 사용 ③ 준거참조검사의 성격이므로 신뢰도로서 의사결정의 일치도를 보일 것	① 재검사법 (기간: 일주일?) ② 반분법.	① 재검사법 (3개월 간격) ② 반분법	반분법
측정의 표준오차보고	X	X	X	X
신뢰도값	.77~.82	.47~.90 (α계수는 낮고 반분신뢰도와 재검사신뢰도가 높았음)	.71~.90	.60~.88

표2 계속

타당도의 증거	미국의MBTI	Cattel16PF		Personal Orientation Inventory
1. 내용타당도				
2. 준거타당도	X	X		X
3. 구성개념타당도	① 영어판과 한국어판간의 수렴타당도 ② 문항비중의 예언도를 변별타당도로 간주하고 있음. ③ 예언도의 개념 설명없음 ④ Jung유형검사(JTS)와 비교(수렴타당도)	① 어느 척도의 점수가 어느 집단에서 높고 어느 집단에서는 낮음을 보임. ② 소검사간 상관계수 제시 ③ 요인분석으로 2차 요인추출(수렴 및 변별타당도의 소극적 노력)	① "논리적 타당도에 대한 사용자의 식견과 비판에 맡기고 생략한다"(p.14)라고 하였음 ② "타당도계수는 지금 진행중인 연구가 끌나는대로 아래의 표에 보완 제시키로 한다"(p.15)고만 하였음.	① 소검사간 상관 ② 시간효율성-비효율성간 상관과 외향성-내향성간 상관은 -.95 안팎임. 상관되는 각쌍의 척도들이 두 척도로 변별되지 않음을 의미.
검사실시·해석하는 사람의 자격	기초 및 보수교육받은 사람에게만 검사지 판매	임상장면에서도 용도를 추천하고 있으므로 그 경우 사용자의 자격 명시 필요.	임상척도에 한해서 "상담교사, 임상심리 전문가…"등으로 명시하였음	규정불필요
검사의 과학적 기초에 대한 증거수집	융의 심리유형론은 최근의 성격이론이 아님.	문항분석에서 척도별 문항-총점상관이 0.2가 안되는 문항이 48개. 전체의 30%	문항작성의 정확한 근거가 빈약('광범위한 참고'라는 진술은 부족	국내에서의 타당도연구가 계속 있어야 할.
논리의 철저한 설명 및 용도를 지지하는 자료와 연구	절수계산공식에서 x2-1 또는 x2+1의 해설필요. 문항점수를 0, 1, 또는 2로 매긴 논리제공 필요.	임상장면의 용도를 추천하므로 임상고객들에 대해 특별히 높은 정보를 제공하는 문항을 보고할 것.	①논리의 설명은 철저하지 않음. ②활용영역은 추천하는 면서 막상 그에 대한 자료·연구는 제시하지 않음	①논리의 설명이 부족 ②용도의 추천을 지지하는 국외에서의 자료나 연구라도 제공했어야 함.

지 않은 것 같다. 성격진단검사에서는 모든 타당도에 대해서 언급하고자 의도는 보이고 있으나 실제로 그 값을 보이지는 못하고 있다.

ㅊ. 준거타당도

성격검사의 성질상 내용타당도와 구성타당도를 보이는 것은 가능하지만 준거타당도를 보이는 것은 쉽지 않다. 성격검사에 대한 준거타당도를 보이려면 이론적으로 예측되는 행동적 준거를 설정하는 것이 우선순서인데 이것은 산업장면에서와 같이 특정의 바람직한 행동이 있는 곳이라면 가능하다. 그 이외의 일상생활의 장면에서 어떤 성격에 대한 준거행동을 설정하기란 쉽지 않다. 그래서 그런지 어떤 검사도 이 타당도를 능동적으로 추구하지는 않았다.

ㅋ. 구성개념타당도

대체로 많이 사용되고 있는 '소검사간 상관의 제시'를 성인용성격검사와 자아실현검사에서 보이고 있다. 그러나 아쉬운 것은 요인분석을 활용하여 어떤 척도들이 어떤 상위요인으로 묶이고 어떤 상위요인으로는 묶이지 않음을 보이므로서 수렴성 및 변별성을 보일 수가 있는데 이러한 분석방법이 아직은 보편화 되어있지 않은 것으로 보인다. 성인용성격검사에서, 1차요인에서 2차요인을 추출함은 그와 같은 활용의 예로 볼 수 있다. 그러나 요인분석을 이 방식으로 사용하는 것은 구성개념타당도를 세우는 소극적 노력에 지나지 않는다. 성인용성격검사에서 척도의 점수들이 집단별로 차이가 있을 것을 추론한 후 실제 자료가 그러한 추론을 지지함을 보인 것은 타당도를 세우는 적극적인 접근의 예이다. 물론 이것 하나로 충분하지는 않다.

MBTI에서 한국판을 영어판에 상관시키는 구성개념타당도 확보의 노력을 보였다. 그러나

문항비중의 예언도를 변별타당도로 간주하는 것은 예언도에 대한 개념설명이 없으므로 이해가 되지 않는다. 예언도가 무슨 개념이며 어떻게 변별의 개념과 연결되는지 교본에 좀 더 해설이 필요하다.

자아실현검사에서 (시간의 효율성과 비효율성) 그리고 (외향성과 내향성)의 괄호안 변수들을 각각 낱개의 척도로 취급하면서 척도간 상관계수를 구했는데 그 값이 -.95를 전후해서 있으므로 효율성/비효율성은 하나의 연속선상에 있는 양극으로 볼 수가 있고 외향성/내향성도 마찬가지이다. 따라서 각각 별개의 척도로 취급하는 것은 정확한 취급이 아닌 것으로 보인다.

어떤 성격진단검사에서는 타당도에 대한 개발자의 의무를 면제받는 듯한 진술을 사용하고 있는데, 이것은 의도와 달리 무책임하다는 느낌을 줄 수가 있다.

ㅌ. 검사실시·해석하는 사람의 자격

성인용 성격검사는 임상장면에서도 사용을 추천하고 있는데 그 경우 사용자의 자격이 교본에 명시되어야 할 것이다.

ㅍ. 검사의 과학적 기초에 대한 증거

MBTI가 50년전의 이론인 융의 "심리유형이론의 타당성을 확증하였고, 이러한 확증에 따라 성격유형을 잡아낼 수 있는 타당한 질문지를 작성"(김정택, 심혜숙, 1991, p.16)하였다는 것은 오늘날 성격이론이 시사하는 것과 상반되고 있다. 오늘날 성격이론의 발전된 모습은 성격을 유형으로서 보다는 연속적 차원(trait)으로서 보고 있다. 따라서 외향성/내향성도 한 차원의 양극이지 MBTI에서처럼 두 차원의 척도로 취급할 성질은 아니라고 본다.

성인용성격검사 역시 문항분석의 부분을 보면 각 척도별 문항-총점상관이 너무 작은 문항들이 많다. 이 상관은 변별도로 쓰이는 것인데 0.2이하 일 때는 그런 문항은 변별도가 너무 적은 것으로 보고 수정하거나 다시 개발할 것을 권하게 된다. 이 검사에서는 이런 문항이 전체의 30%가 되고 있으니 빈약한 문항이 너무 많은 셈이다.

성격진단검사 역시 다른 척도들(예:MMPI)을 광범위하게 참고하였다고 진술하고 있으나 보다 정확한 근거를 제시해야 할 것이다.

ㅎ. 논리의 철저한 설명, 용도지지하는 자료·연구

MBTI의 경우 유형별 점수계산공식에서 "(큰숫자-적은숫자)x2±1"의 공식이 쓰이고 있는데 x2와 ±1의 근거가 되는 연구를 소개하고 그 논리를 철저히 해설하는 것이 검사의 논리를 제시하는데 도움을 줄 것이다. 또한 문항의 점수도 경우에 따라 0, 1, 또는 2점으로 주고 있는 논리가 제시되어야 할 것이다.

성인용성격검사에서도 그 검사에 대한 임상장면의 용도를 추천한다면 임상내담자들에게 특별히 높은 정보를 제공하는 문항들만을 사용하므로서 일반 정상인들에게 해당되는 문항들을 굳이 내담자들에게 적용시키는 일을 방지할 수가 있다. 그러기 위해서는 보다 고급적인 검사이론의 틀 속에서, 임상내담자의 속성수준에서 각 문항이 제공하는 정보를 계산하고 정보가치가 큰 문항들만을 선정하도록 해야 할 것이다.

적성검사

(1) 세부검토

여기서 검토되는 적성검사는 다음과 같다: 적성종합검사(정범모, 김호권, 1992), 적성진단검사(이상로, 김경린, 1974), GATB 직업적성검사(박수병, 1967)

(2) 적성검사의 요약

ㄱ. 규준제작 이후 20년이상 경과한 것이 두 개나 된다.

ㄴ. 규준의 수효측면에서 볼 때, GATB는 중 1, 2, 3, 고교생의 4가지 규준이 있다. 그러나 적어도 남자/여자의 구분만큼은 필요하지 않을까 생각된다. 즉 사용자들의 다양한 상황을 고려하지 않고 적은 수의 규준밖에 없는 경우가 아닌가 한다.

표3. 적성검사의 검토

	적성종합검사	적성진단검사	GATB
제작자	정범모, 김호권	이상로, 김경린	박수병
발행자	코리안테스팅센타	중앙적성출판사	중앙적성출판사
규준제작년도 (자료수집기간)	? (?)	1974 (?)	1967 (66.11-66.12)
규준제작후 경과	?	22년	29년
규준	12개	7개	4개
표집방식	전국 고교 학년별, 도시/지방, 남자/여자별 층화표집	① 대학생: 지역, 학교, 학과, 학년, 성별로 층화표집 ② 일반인: 직종·직위별로 층화표집	X
총문항수	160개. (이 글에서의 목적상 어휘력은 2문항, '수·공능력'은 1문항으로 계산하였음)	235개	295개 (문제에서의 번호와 이 글의 목적상 계산된 문항수와는 다름)
규준집단의 크기	1083~2334	532~986	1517~2469
관찰-문항비율	모두 5배가 넘음	2배는 넘지만 5배가 되는 표본은 없음	모두 5배가 넘음
신뢰도 보고구분	① 소검사별 보고 ② 남자/여자로만 구분해서 보고 ③ 도시/지방별로 신뢰도의 차이는 없을지?	① 소검사별 보고 ② 인구집단별 신뢰도 차이 없을지?	① 소검사별 보고 ② 고등학생과 중학생 사이에 신뢰도 차이는 없을지?
신뢰도 계산방식	반분법 (지각속도, 어휘력, '수공능력'외에 속도검사의 성격이 없다고 주장)	KR-20 반분법	재검사법(50일 간격)
측정의 표준오차보고	X	X	X
신뢰도값	.81~.95	.51~.76	.68~.93
타당도증거			
1. 내용타당도	표지에는 고·대·일반용이라고 했는데 일반인용의 규준은 없음	X	X
2. 준거타당도	X	X	각 척도를 학업성적과상관
3 구성개념타당도	① 소검사간 상관 ② 각 소검사로부터 요인추출	① 요인분석: 4가지 적정 요인으로 수렴. ② 이 검사의 전신인 '고등학생-성인용 적성진단검사'와 상관은 일종의 수렴타당도 ③ 지능과의 상관있음은 역시 수렴타당도	X
검사실시자 자격	규정불필요	규정불필요	규정불필요
검사의과학적 기초	추후의 후속연구를 통한 보강이 안되고 있음	후속연구의 보고가 없음	후속연구의 보고가 없음
논리의 철저한 설명, 용도지지하는 자료 및 연구제공	X	X	X

ㄷ. 관찰-문항비율은 적성진단검사를 빼고는 대체로 5배가 넘는다. 적성진단검사의 경우 표본크기가 좀 더 커야 할 것이다.

ㄹ. 신뢰도 보고시에 척도별로 보고하는 것은 당연하지만 인구집단간에 신뢰도의 현저한 차이가 기대되는 데도 보고하지 않는 것은 적어도 그에 대한 논의가 제시되어야 할 점이다.

ㅁ. 신뢰도 계산방식

적성검사는 본질적으로 시간제한이 주어지므로 재검사법이 선호될 것인데 GATB에서만 이 방법이 채택되고 있다.

ㅂ. 측정의 표준오차

어떤 적성검사도 이것을 보고하지 않고 있다.

ㅅ. 신뢰도의 값

Mehrens와 Lehmann(1980)의 기준에서 볼 때 세 검사중 적성진단검사의 신뢰도가 가장 부족하다.

ㅇ. 내용타당도

적성종합검사의 경우 표지에서 '일반용'에 쓸 수 있다고 하고 있는데 실제로 규준은 대학생까지만 있다. 내용타당도에 크게 의문이 가는 경우이다. 전반적으로 내용타당도에 대한 증거의 제공은 약하다고 할 수 있다.

ㅈ. 준거타당도

적성검사 성적과 실제 직업에서의 성공여부, 학업 등과 상관을 보아야 할 것인데 GATB만이 각 척도를 학업성적과 관련시켜 보았다. 그런데 이 때도 서울의 3개 고교(경기고, 이화여고, 상명여고) 169명에 대해서만 조사했으므로 일반화가능성이 크게 제한된다.

ㅊ. 구성개념타당도

GATB는 전혀 노력을 보이지 않고 있다. 적성종합검사와 적성진단검사에서 소검사간 상관과 요인분석을 통해서 소극적 접근을 하고 있다. 그러나 적성진단검사에서 전신인 '고등학생-성인용 적성진단검사'와의 상관, 그리고 지능과의 상관을 구한 것은 수렴타당도를 위한 적극적 접근을 보여주고 있다.

ㅋ. 공통의 문제점

적성검사들이 갖는 공통의 문제점은, 소검사에서의 점수가 몇개의 적성요인에서의 점수가 되며 응답자가 어떤 직업군을 추천받으려면 각 요인에서의 최소의 기준점수가 어떻게 산정되어야 하는 가이다. 적성종합검사는 기준점수를 80%의 확률에 의거했다고 한다. 즉 한 직업분야에 들어가서 성공한 학생들중 80%가 이 기준점수보다 높은 점수를 받았다는 것이다. 그러나 GATB는 기준점수에 대한 논리의 설명이 전혀 없다. 즉 적성검사들에서 이와 같은 류의 논리에 대한 철저한 설명 및 적성검사점수와 실제 직업분야에서 종사자들에게서 관찰되는 특성과의 관련을 또는 일치되는 정도를 사용자들에게 제시할 수 있어야 한다(표준 3.6).

임상검사

(1) 세부검토

여기서 검토된 임상검사는 다음과 같다: 다면적 인성검사(김영환등, 1989), 간이정신진단검사(김광일, 김재환, 원호택, 1984). 전자는 MMPI, 후자는 SCL로 부르기로 한다.

(2) 임상검사의 요약

ㄱ. 규준은 대체로 아주 오래되지는 않았다.

ㄴ. MMPI의 경우 비교적 많은 종류의 규준이 있어서 사용자가 자신의 상황에 맞는 규준을 참조할 수 있는 여지가 크다.

ㄷ. 표집방식은 SCL의 경우 분명하게 기술되어 있지는 않다.

ㄹ. 관찰-문항비율

SCL의 경우 비교적 충분한 표본크기지만 MMPI는 문항수에 비해 표본이 작다고 할 수 있다. 또는 문항의 수효가 압도적으로 많아서 관찰-문항비율이 작을 수 밖에 없다고 볼 수도 있다. 따라서 MMPI는 동일한 정도의 효능을 갖는 축소형검사 또는 내담자 수준에 맞게 재단된 검사가 적극 필요할 것이다. 이와 같은 방향은 문항반응이론을 사용할 경우 용이할 것

표4. 임상검사의 검토

	MMPI	SCL
제작자	김영환등	김광일등
발행자	한국가이던스	중앙적성
규준제작연도 (자료수집기간)	1989 (88.8-89.1)	1984 (?)
규준제작후 경과	7년	12년
규준	22개	10개
표집방식	학생, 일반집단에 대하여 성별, 연령, 학력, 거주지를 단위로 전국에서 층화표집	편의표집(?)
총문항수	566개	90개
규준집단의 크기	375~931	247~1013
관찰-문항비율	모두 2배 미만	9개의 집단이 모두 5배 이상. 1개집단도 2배 이상은 된다.
신뢰도 보고구분	①소집단별 보고 ②응답자들을 합하여 한 가지로 계산 ③과연 22개의 규준집단에서의 신뢰도가 모두 같을까?	① 소집단별 보고 ② 응답자들에 대해서 한가지로 계산하였음 ③ 규준집단간에 신뢰도 차이가 없을까?
신뢰도 계산방식	반분신뢰도, α계수, 재검사(2주후)	재검사(1주일 후), α계수
측정의 표준오차보고	×	×
신뢰도값	반분 신뢰도:.48~.86 α계수:.52~.84 재검사신뢰도:.52~.89	재검사신뢰도:.73~.83 α계수:.68~.89
타당도 증거 1. 내용타당도 2. 준거타당도 3. 구성개념타당도	미국 MMPI × ① 소검사간 상관 ②과거의 MMPI(1963년판)를 사용한(일부문항) 점수와 높은상관이 있음을 보일 것.	미국 SCL-90-R × ① 요인의 불변성계수: 같은 요인을 두 개의 상이한 집단에서 쟀을 때 일치하는 정도이므로 일종의 수렴타당도. ② 한국판에 대해서 타당도의 증거가 별로 없음을 정직하게 밝히고 있음(표준 1,2)
검사실시자 자격	명시적 진술이 필요한데 기술되지 않았음.	명시적 진술이 필요한데 기술되지 않았음.
검사의 과학적 기초	명시적으로 기술되어 있지는 않음	충분히 기술되어 있지 않음.
논리의 설명, 용도 지지하는 자료·연구	충분치 않음	충분치 않음

이다. 국내에서도 MMPI에 문항반응이론의 적용을 시도한 예가 있다(예: 신석기, 1987).

ㅁ. 신뢰도 보고구분

척도별로 보고하는 것은 잘 지켜지고 있으나 규준집단간에 신뢰도의 차이가 현저히 날 수 있는 가능성을 배제할 수 있는 논의가 없이 한 가지로 신뢰도를 계산하고 있음은 사용자에게 부정확한 정보를 줄 우려가 있다.

ㅂ. 신뢰도 계산방식

비교적 여러가지 방식으로 계산하고 있다.

ㅅ. 측정의 표준오차

전혀 보고되지 않고 있다. 아직은 이 개념

이 임상장면에서 유용하게 쓰이지 못함을 반증하는 것으로 본다.

ㅇ. 신뢰도값

특히 Mehrens와 Lehmann(1980)의 기준에서 볼 때 모두가 바람직한 수준의 신뢰도는 아니다. MMPI의 경우 척도에 따라서는 신뢰도가 낮은 것들이 좀 있다.

ㅈ. 내용타당도

모두 미국의 원본척도에 의존하고 있으나 미국의 임상집단과 한국의 임상집단에 다른 점이 있다면 문항들의 내용대표성도 달라질 수 있지 않을까 생각된다.

ㅊ. 준거타당도

임상척도일수록 이 준거타당도가 아주 중요하다. 예로서 이 검사의 점수가 다른 사람에 의해서 이미 진단받은 임상내담자의 상황과 일치되는 정도를 보고한다면 좋은 준거타당도의 증거가 될 것이다.

ㅋ. 구성개념타당도

임상심리학의 이론체계속에서 관련된 변수들간의 관계를 이론적으로 가설화하여 검증한 것이 제시되면 구성개념타당도의 좋은 증거가 될 것이다. MMPI의 경우 적어도 구판(1963년판)의 검사들이 나타내는 것과 상당정도 수렴함을 보여야 할 것이다.

ㅌ. 검사실시자의 자격

이 두 검사는 명백히 임상장면에 쓰이는 검사이므로 그 사용자의 자격요건을 제한하여 피검자가 무자격한 검사의 실시 및 해석의 피해를 입지 않도록 해야 할 것이다.

ㅍ. 검사의 과학적 기초

MMPI나 SCL에 대한 국내에서의 지속적인 타당화연구를 통해 그 과학적 기초가 튼튼함을 보여야 할 것이다. 즉 이 부분을 보다 명시적으로, 충분히 기술하므로서 검사의 품질을 높일 수가 있을 것이다.

ㅎ. 논리의 철저한 설명, 용도를 지지하는 자료·연구의 제시

이 검사들의 논리 및 용도를 지지하는 자료와 연구가 충분히 제시되어야 할 것이다. MMPI의 경우 논리 및 용도는 어느 정도 설명되어 있으나 용도를 지지하는 자료와 연구를 증거자료와 함께 제시하는데 아주 인색하다. SCL 역시 마찬가지이다.

검사의 타당도에 대한 강조 필요

이제껏 기술적 표준의 측면에서 국내의 검사도구들의 품질을 알아보았다. 품질에서 핵심이 되는 개념이 타당도와 신뢰도임에는 이의가 없다. 그러나 국내의 각 검사들에 대한 검토에서 본 바와 같이 신뢰도의 이해, 계산, 및 보고는 대체로 보편화되어 있다. 물론 바람직한 수준이 되려면 아직은 미흡하다. 또한 검사의 종류가 규준참조검사가 아닌 준거참조검사일 경우 종래의 방식인 평행성원리(parallelism)에 근거한 세 가지 방식 -- 검사·재검사법, 동형검사법, 내적일관성법 -- 이 아닌 다른 개념의 신뢰도가 필요할 것이라는 것은 강조될 필요가 있다. 신뢰도 개념이 비교적 익숙한데 비해 타당도의 적극적 접근은 상대적으로 덜 보편화되어 있다. 따라서 지금부터 타당도에 대한 새로운 추세와 타당도의 일반화에 대하여 논의하기로 한다. 이 글의 주된 목적이 현재 우리나라의 표준화검사의 기술적 품질의 현주소에 대한 고찰이었으나 그 결과로 타당도에 대한 개념 및 타당도 연구의 보편화가 가장 시급하다고 판단되어 간단히 타당도에 대한 논의를 첨가하기로 한다.

검사의 타당도에 대한 새로운 추세

새로운 추세를 소개하기 위해서 역사적 관점에서 고찰하기로 한다. 우선 검사의 타당도에 대한 정의가 어떻게 변해 왔나를 보기로 한다. 1940년대에 검사의 타당도를 보이는 책임은 전적으로, 검사를 개발 또는 그 검사를 어떤 목적에 쓸 것을 추천하는 사람에게만 있었다(Angoff, 1988). 이 당시 검사의 타당도는 '그 검사가, 무엇인가 측정하도록 되어 있는 바의

것을 측정하는지'의 정도로 정의되었다. 예컨대 검사가 인력선발을 위한 것이면 검사점수와, 나중에 고용된 후의 행동(예: 업무성과)간의 상관계수를 구해서 예측타당도의 계수로 제시하게 된다.

1950년대부터 타당도의 정의는 '검사가, 그걸 사용한 목적을(예: 인력선발, 기능숙달판정) 이루는 가를 보이는 것'으로 변해간다. 이 새로운 정의는 검사의 타당도를 보이는 책임의 절반을 검사의 사용자에게 지우게 된다. 이 추세는 계속 발전하여 요즈음의 검사의 타당도에 대한 정의는 "검사 자체에 대한 것 또는 검사점수에 대한 것이라기보다는 검사의 사용자가 검사점수를 보고 내리는 해석 및 유추와 그 유추에 근거한 의사결정 및 행동의 타당도를 보이는 것"(Angoff, 1988, p.24)으로 발전해 왔다. 이 개념에는 물론 검사의 개발에 대한 문제도 포함한다. 그러나 이제는 개발만이 문제가 아니고 검사실시, 거기서 나온 정보의 해석 및 의사결정까지가 검사의 타당도를 보일 때 논의되는 범위가 된다.

단순히 일반사회에서 검사의 한계에 대한 이해가 없는 문외한들에게는 '검사가 측정하고자 한 것을 측정하는가'에 대한 측정타당도는 경우에 따라 실망스럽고 검사 무용론으로 몰고갈 정도로 낮다. 예로서 지능검사가 재는 학문적 지능은 학교성적과 가지는 상관계수의 범위는 .4~.7 이지만, 학교가 아닌 실생활에서의 업무수행과는 .2~.25 정도의 관계밖에 없다 (Wigdor & Garner 1982). 그러나 그렇게 측정의 타당도가 낮은 검사라 해도 상황에 따라서는 그 검사를 사용하지 않았을 경우의 의사결정의 능률에 비해 수십 %씩의 증가된 의사결정능률을 보인다. 여기서 능률이라 함은 내려진 의사결정중에서 올바른 결정의 비율을 말한다. 물론 이와같은 경우의 타당도는 검사점수를 가지고 계산한 것이 아니라 의사결정을 가지고 계산한 것이므로 의사결정의 타당도라고 할 수 있다. 산업장면에서 측정의 타당도를 이야기 하면 관리자를 실망시키는 바가 많지만 의사결정의 타당도를 이야기 하면 검사의 긍정적인 측면을 잘 전달하는 기회가 될 수 있다. 따라서 품질 또는 양호도를 고려할 때 단순히 측정의 타당도만이 아니라 의사결정의 타당도를 적극적으로 고려해야 할 것이다.

또한 오늘날의 추세에서는 개발자와 사용자가 함께 타당도에 대한 책임을 지게 되며 어느 한 쪽 혼자서도 타당도에 대한 책임을 완벽하게 담당할 수 없다. 따라서 사용자들도 적정한 소양을 가져야 검사의 타당도가 현실적으로 이해될 수 있다. 현재 미국 심리학회에서는 검사를 실시 또는 검사점수를 해석하는 사람들에 대한 자격요건을 연구하는 위원회가 임명되어 있다.

다음은 검사 타당도의 세부 측면에 대한 변천을 보기로 한다. 1940년대에는 예측타당도를 중심으로 검사의 타당도을 보이는 추세였다. 1950년대에는 내용타당도를 보이는 것이 또 하나의 측면으로 추가된다. 1954년의 검사표준서는 타당도의 종류로서 종래의 예측타당도(predictive validity), 현측타당도(concurrent validity; 일치타당도, 공인타당도)[2] 및 내용타당도외에 구성개념타당도(construct validity)를 추가한다. 이 추세는 1950년대 이전에 주로 계량분석(psychometrics) 중심의 타당도연구가 지배적이던 것이 1950년대 이후에는 내용적 이론의 중요성이 점점 강조되는 추세에 있음을 보여 준다. 물론 계량분석중심의 타당도연구는 그 이전의 탁상공론(armchair theorizing)에 대한 반발로서 나온 것이었다.

1966년과 1974년의 검사표준서에서는 예측

---

2) concurrent validity는 여러 가지로 번역된다. "현재"시점에서 확보되어 있는 준거변수에 비추어 검사의 타당도가 논의되므로 현측타당도로 번역하였다. 미래시점에서 얻어지는 준거에 비추어 논의되는 타당도를 예측타당도라 하고, 과거시점에서 확보된 준거에 비추어 논의되는 타당도는 과측(postdictive)타당도라고 하는 맥락에서 concurrent validity를 현측타당도로 번역한 것이다.

타당도와 현측타당도를 준거타당도의 두 측면으로 묶었다. 따라서 1970년대에는 검사의 개발자나 사용자는 준거타당도, 내용타당도, 또는 구성개념타당도 중 어느 하나, 두 가지, 또는 세 가지 종류를 보임으로서 검사가 타당화된다고 생각했다. 이러한 제한적인 견해는 미국의 경우 고용평등법(EEO Law)에서 지금껏 그대로 반영되어 있다. 즉 기업의 입장에서 입사시험의 타당도를 보이기 위한 전략으로서 세 가지 중 하나만 보여도 되는 것으로 되어 있다. 그러나 80년대에는 이들 세 가지 종류를 서로 독립된 것으로 보지않고 구성개념타당도라는 우산아래 하나의 주제로 묶어서 보자는 주장이 제기되었다(예: Messick, 1988). Jones와 Appelbaum(1989)도 이 추세에 동의하기는 하나, 예측타당도와 내용타당도가 교육및 산업분야에서는 계속 개별적으로 특별한 주목을 받을 것임을 강조한다. 교육및 산업분야의 검사에서, 각 문항들은 검사에 관련된 전체 영역에서 뽑힌 것들이어야 하며(내용타당도) 바람직한 결과(예: 좋은 성적, 모범적 품행의 학생, 유능한 사원)를 예측할 수 있는가(예측타당도) 하는 것이 크게 문제가 되기 때문이다. 그러나 구성개념에 강조를 두는 현재의 추세가 시사하는 바는 분명히 수용되어야 할 것이다.

또한 80년대에 주목받기 시작한 준거참조 검사에서는 다른 무엇보다도 문항들이 전집(domain)에서 나온 것인가에 대한 관심을 가지다 보니 내용타당도가 부각되고 있다(Popham & Husek 1969). Messick(1980)은 내용타당도를 내용대표성으로 용어를 바꾸어 부를것을 주장한다. 그 이유는 '타당도'라는 것은 반드시 검사점수를 가지고 정의되는 것인데 이제껏 보아온 "content validity"의 노력은 검사점수가 아닌 검사도구(test instrument, test forms)의 그럴듯함에 국한된 것이므로 굳이 "validity"란 말을 여기에 쓸 필요가 없다는 주장이다. 내용타당도는 검사를 응답자들에게 실시한 자료에 의해 판단하는 개념이 아니므로 계량화 하기는 어렵지만 꾸준한 노력은(예: Lawshe, 1975) 계속 이어지고 있다. 최근의 Applied Psychological Measurement를 보면 문항을 실시하기 전에 문항간 유사성에 대한 자료를 수집하여 다차원척도법(Multidimensional Scaling)이나 군집분석(Cluster Analysis)를 사용하여 내용타당도를 보려는 논문들이 가끔 눈에 뛴다. 따라서 학과목의 시험처럼 논리적 표집(logical sampling)의 타당도를 통해서 내용타당도를 보일 수가 있는 경우와는 다른 입장에 있는 성격·태도·의견 등의 개념에 대해서도 보다 계량적인 접근의 가능성이 높아가고 있다. 또한 구성개념타당도라는 큰 범주에 "준거타당도"도 한 부분으로 들어가므로 predictive validity는 predictive 'utility'로 concurrent validity는 disgnostic 'utility'로 바꿔야 한다는 것이 Messick의 주장이다. 검사관행에서 구성개념의 확보가 그만큼 중요함을 나타내는 정도로 저자는 이해한다.

검사타당도에 대한 연구를 위해 Cronbach과 Meehl(1955)이 이미 주장했듯이 우리는 재고자하는 구성개념(construct)에 대한 검토를 해야하며 그 개념에 대한 이론과 검사점수간에 서로 조화시키는 연구를 꾸준히 해야 한다. 이것을 Embrestson(1983)은 법논리같은 연결망(nomothetic span) 이라고 부른다. 이는 측정되는 개념의 이론적 내용은 검사결과에 대한 해석을 위해 지침을 제공하고 또 경우에 따라서는 어떤 자료가 더 필요한 가의 방향을 제시하는 한편, 검사결과는 그 이론적 내용 자체를 검토하고 수정 또는 기각하는데 사용됨을 의미한다. 따라서 구성개념타당도는 구성개념에 대한 이론적 합당함을 보이고서 한번으로 끝나는 절차가(procedure) 아니고, 경험자료에 근거하여 계속적으로 진행되는 과정(process)으로 보아야 한다는 것이다. 그래서 Angoff(1988)는 오늘날의 시대는 경험적 이론(data- based theory)에 의해 구성개념타당도를 보이는 시대라고 하고 있다. 그러나 타당도를 보이기 위한 방법에는 단 한 가지가 아닌 많은 전략이 있으며 반드시 계량적인 자료만이 타당도의 증거가

될 수 있는 것은 아니라는 것을 명심해야 할 것이다. 검사이론에도 포스트모더니즘의 바람이 분다면 그것은 계량일변도에 질적판단의 중요성을 부각시키는 움직임일 것이다. 예로서 타당도의 증거를 꼭 계량적이 아닌 질적인 것으로 할 수도 있다는 의미이다.

한편 60년대에 부활한 인지심리학의 바람이 이제는 검사이론에도 상당한 영향을 미치고 있다. 예로서 Embretson(1983, 1985)은 과거의 구성개념타당화의 노력은 단지 관련 변수들과의 법논리같은 연결망만을 확보하려는 행동이었다고 지적하면서 이제는 '구성개념의 표상'을 타당화노력에 포함시킬 것을 주장한다. '구성개념의 표상'이란 문항반응에 정신과정, 문제풀이 전략, 및 지식의 축적량 등이 포함됨을 의미한다. 즉 검사를 제작할 때 '구성개념의 표상'이 안된 상태에서는 아무리 계량자료에 근거한 논리적 연결망을 제시해도 구성개념의 존재를 입증하지 못한다는 것이다.

타당도의 일반화

똑 같은 검사라도 어떤 환경에서 실시하느냐에 따라 타당도계수가 크게 다른 값이 나오는데 그 이유는 여러 가지가 있다. 예컨대, Schmidt와 Hunter(1977)는 인력 선발검사의 예측타당도가 크게 상황에 의존하는 이유를 다음의 이유로 돌린다: 표집오차(sampling error), 직무별로 각 준거변수(criterion)의 신뢰도가 다를 수 있고, 응답자들에 다양함이 없음(range restriction) 등. 또한 대부분의 타당도계수가 상관계수로 표시되므로, 상관계수의 값을 편파시킬 수 있는 다음의 요인들은 타당도계수의 값들이 서로 다르게 나오는 이유가 될 수 있다고 본다: ① 검사점수와 준거점수간에 비직선적관계(non-linearity)가 있어도 상관계수는 작은 값이 된다, ② 예외적인 점수들(outliers), ③ 검사점수 또는 준거점수의 폭의 왜곡(range enhancement and restriction).

타당도의 계수로서 보통은 상관계수 또는 처치효과의 크기(effect size: 실험의 경우)가 사용된다. 예컨대 검사의 개발자가 예측타당도를 보고할 경우, 검사도구에 응답한 사람들의 점수가 어느 정도 시간(세월)이 지난 후 학업에서 또는 직업에서의 성과(performance)점수에 얼마나 상관이 되는 가를 예측타당도의 값으로 보고한다. 이 때 이 값이 과연 얼마나 일반화될 수 있는지는 단 한번의 연구에서 나온 타당도의 값을 통해서 결론내리기가 어렵다. 그래서 각 독립된 연구에서 제시되는 타당도의 값을 종합해서 종합적인 타당도계수를 알아보는 것이 타당도의 일반화를 위한 노력이다.

검사표준서에서는 타당도의 일반화(validity generalization)를 다음과 같이 정의하고 있다:

"하나 또는 그 이상의 특정의 상황에서 얻어진 타당도 증거를 다른 유사한 상황등에 적용하는 것. 이 때 동시추정(simultaneous estimation), 상위분석(meta-analysis), 또는 합성적 타당화(synthetic validation)의 논의에 기초하여 일반화 연구를 진행한다" (이순묵·이봉건, 1995, p. 193).

타당도의 일반화에 대해서 검사표준서에서 언급하고 있는 세 가지 접근법을 간단히 해설하면 아래와 같다.

① 동시추정: 검사의 동시적용

일단 검사가 개발되면 두 개 이상의 상황에서 독립적으로 동시에 적용시켜 그 타당도의 정도가 비슷한지 비교해 본다.

② 상위분석: 개별연구의 종합

타당도계수가 독립적으로 여러 곳 또는 여러 경우에 구해졌을 때 그 결과가 반드시 일치하지는 않을 경우도 많다. 이 때 그 결과를 통계적으로 종합해서 타당도에 대한 가설을 세우고 그것을 검증하는 방법이다. Hunter와 Schmidt(1990)의 책이 좋은 참고가 된다.

③ 합성적 타당화: 요소검사의 합성

산업심리학에서의 예를 가지고 이 개념을 설명하기로 한다. 산업장에서 타당도의 일반화는 보통은 어떤 직무를 대상으로 해서 이루어진다. Mossholder와 Arvey(1984)는 어떤 직무를

위한 인력선발에서 검사를 사용하여 이상적인 도움을 얻는 경우는 다음과 같은 조건에서라고 한다:

ㄱ. 그 직무를 수행함에 필요한 어떤 속성들이 분명히 정의되어 있고

ㄴ. 이들 속성들을 재는데 타당도있는 검사들이 이용가능하며

ㄷ. 검사점수들과 직무수행(준거점수) 사이에 잘 맞는 관계가 있을 때이다.

이런 조건하에서 개발된 검사가 여러 상황에 일반화 될 수 있을 때 일반타당성이 있다고 한다. 그러나 이런 조건들은 항상 가능한 것은 아니다. 즉 어떤 직무에 몇 명 안되는 직원밖에 없을 경우 수집할 자료조차 넉넉치 않은 것이다. 또한 각 직무별로 타당도있는 검사체계를 만든다는 것은 많은 경제적 지출을 필요로 한다. 이런 경우 어떤 직무에 대해서 일반타당도있는 검사체계를 세우려고 하는 것보다는, 요소검사의 합성을 통해 검사체계를 세우는 것이 더 경제적인 동시에, 중소기업처럼 각 직무에 소수의 인원만이 있는 경우엔 현실적인 방식이다. 요소검사의 합성은 직무를 그 구성요소별로 분석해서, 각 요소에 맞는 검사가 가지는 타당도를 결정한 후 이들 타당도를 한 덩어리로 합성하는 방식의 타당도 추구를 말한다; 그래서 synthetic validity를 "요소검사의 합성"이라고 번역한다.

사실 "synthetic validity"는 또 하나의 타당도정의가 아니고 이미 여러 가지 직무구성요소에 대해 타당도있는 검사들이 있을 때, 몇 가지 구성요소의 집합으로 정의되는 하나의 직무에 대한 집합적 타당도(test-battery validity)를 유추하는 논리적 과정이다(Guion, 1965). 예컨대 A, B, C 및 D라고 하는 검사가 각각 직무요소 1, 2, 3, 및 4에 대한 타당도있는 검사이고 어떤 직무가 요소 1과 3의 집합으로 정의될 경우 검사 A와 검사 C를 합성하면 이 직무에 적합한 사람을 선발하는 타당도있는 검사체계가 된다는 것이다. 이러한 방식으로 검사체계의 타당도를 세우는 과정이 곧 요소검사의 합성이다. 국내에서 조직의 규모에 따라서는 자체적인 인구를 가지고 어떤 검사의 타당도를 시도할 수 없는 경우가 많다. 그럼에도 불구하고 인사관리의 목적상 사용하는 검사에 대해 타당도를 보여야 한다면 이 합성적 타당도의 접근이 권고될 수 있다.

참 고 문 헌

김광일, 김재환, 원호택(1984). **간이정신진단검사 실시요강**. 서울: 중앙적성출판부.

김기석(?). **학습습관검사 해설 및 실시요칙(중·고등用)**. 서울: 코리안 테스팅센터.

김영환,김재환,김중술,노명래,신동균,염태호,오상우(1989). **다면적 인성검사실시 요강**. 서울: 한국 가이던스.

김재은,이광자(1983). **자아실현검사 실시요강**. 서울: 중앙적성출판부.

김정택,심혜숙(1991). **MBTI안내서 Myers Briggs Type Indicator Manual**. 서울: 한국심리검사연구소.

김해옥,김난수,임의도,이인수,조석호(1993). **표준화 지능성숙검사 실시요강(국민학교 1, 2, 3학년용)**. 서울: 교육과학사.

박경숙,윤점룡,박효정,박혜정,권기욱(1991). **KEDI-WISC 검사요강**. 서울: 도서출판 특수교육.

박수병(1989). **GATB 직업적성검사 실시요강**. 서울: 중앙적성출판사.

신석기(1987). **1모수 로지스틱모형에 의한 MMPI우울증 척도의 문항양호도 개선**. 부산대학교 교육학과 석사 논문.

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

염태호,박영숙 · 오경자 · 김정규 · 이영호(1992). **K-WAIS 실시요강**. 서울: 한국가이던스.

이상로,김경린(19?). **적성진단검사 실시요강(중학생-성인용)**. 서울: 중앙적성연구소.

이상로,변창진(1991). **직업흥미검사 실시요강 (중학-성인용)**. 서울: 중앙적성출판사.

이상로,변창진 · 진위교(19?). **성격진단검사 실시요강**. 서울: 중앙적성출판사.

이순묵,이봉건(1995). **설문 · 시험 · 검사의 제작 및 사용을 위한 표준**. Standard for Educational ad Psychological Testing(미국심리학회, 1985)의 번역. 서울 : 학지사

이종승(1987). 표준화 심리검사의 양호도 분석. **교육평가연구**, 2, 81-105.

정범모,김호권(1992). **적성종합검사 검사법 요강**. 서울: 코리안 테스팅 쎈타.

정범모,김호권(1993). **일반지능검사**. 서울: 코리안 테스팅 쎈타.

Angof, W. H.(1988). Validity: An Evolving Concept. *In Test Validity.* Edited by Howard Wainer and Herry I. Braun. New Jersey: Lawrence Erlbaum

Cronbach, L. J. & Meehl, P. E.(1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, 52, 281-302

Croker, L. & Algena, J.(1986). *Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart and Winston.

Cronbach, L. J., Glesser, G. C., Nanda, H., & Rajarathan, N.(1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

Ebel, R. L. & Fresbie, D. A.(1991). *Essentials of Educational Measurement.* (5th ed.). New Jersey: Prentice Hall.

Embretson, S. E.(1983). Construct Validity: Construct Representation Versus Nomotheic Span. *Psychological Bulletin,* 93, 179-197.

Embretson, E. E.(1985). *Multicomponent Latent Trait Models for Test Design. In Susan E. Embretson, Test Design : Developement in Psychology and Psychometrics,* New York: Academic Press.

Gorsuch, R. L.(1974). *Factor Analysis.* Philadelphia: W. B. Saunders.

Graham, D. & Bergquist, C.(1975). *An examination of criterion-referenced test characteristics in relation to assumptions about the nature of achievement variables.* Paper presented at the annual meeting of the AERA, Washington.

Guilford, J. P.(1954). *Psychometric Methods* (2nd Ed). New York: McGraw-Hill.

Guion, R. M.(1965). An illustrative study of synthetic validity. *Personnel Psychology,* 18, 49-65

Hambleton, R. K. & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement,* 10, 159-170.

Hunter, J. E. & Schmidt, F. L.(1990). *Method of Meta-Analysis: Correcting error and bias in research findings.* California: Sage Publications.

Jones, L. V. & Appelbaum, M. I. (1989). Psychometric Methods. *Annual Review of Psychology.* 40, 23-43.

Lawshe, C. H(1975). A quantative Approach to Content Validity. *Personnel Psychology,* 28, 563-575

Mehrens, W. A. & Lehmann, I. J. (1980). *Standardized Tests in Education*(3rd Ed.). New York: Holt, Rinehart and Winston.

Messick, S.(1980). Test Validation and the Ethics of Assessment. *American Psychologist,* 35. 1012-1027

Messick, S.(1988). The once and Future Issues of Validity: Assessing the meaning

and consequences of Measurement. *in Test Validity*. Edited by Howard Wainer and Herry Braun. New Jersey : Lawrence Erlbaum.

Mossholder, K. W. & Arvey, R. D.(1984). Synthetic Validity: A conceptual and comparative reviews. *Journal of Applied Psychology,* 69, 322-333.

Nunnally, J. C.(1978). *Psychometric Theory* (2nd Ed.). New York, NY: McGraw-Hill.

Popham, W. J. & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement.* 6, 1-9.

Schmidt, F. L. & Hunter, J. E.(1977). Devolopement of a general solution to the problem of validity generalization. *Journal of Applied Psychology,* 62, 529-540.

Sternberg, R. J.(1985). *Beyond IQ: A triarchic theory of human intelligence.* NY: Cambridge University Press.

Thorndike, R. M., Cummingham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and Evaluation in Psychology and Education.* New York: John Wiley.

Traub, R. E. & Rowley, G. L.(1980). Reliability of Test Scores and Decisions. *Applied Psychological Measurement,* 4, 517-546.

Wigdor, A. K. & Garner, W. R. (Eds.) (1982). *Ability Testing: Uses, consequences, and controvercies.* Washington, D. C.: National Academy Press.

Wolman, B. B. (1989). *Dictionary of Behavioral Science.* San Diego: Academic Press.

韓國心理學會誌
Korean Journal of Psychology
1996. Vol. 15, No. 1, 1-25

Quality of Standardized Psychological Testing : investigation of 13 test instruments

Soonmook Lee

Sung Kyun Kwan University

The quality of psychological testing is defined as the degree to which the testing is standardized. It is emphasized that what is to be standardized is not just the test instrument but the totality that covers logical reasoning of test structure, design, production, administration of instruments, scoring, interpretation of test results, and application for decision making. The whole gamut is called testing. Standards for Educational and Psychological Testing supported by APA, AERA, and MCME(1985) connote that the testing should be standardized. Based on the standards, some test instruments are examined mainly on the degree to which the technical standards are satisfied. If the two major concepts in technical standards are reliability and validity, the former is relatively well understood. However, the latter is not well understood. Especially the active approach of validation and the thought that validation is a continuous process are needed.