한국심리학회지 : 일반
Korean Journal of Psychology
2001. Vol. 20, No. 2, 305-325

영역참조적 검사제작의 논리와 적용

이순묵[*] 김청택[**] 김명소[***] 설현수[****]
유태용[*****] 이도형[*****] 임대열[******]

성균관대 BK21 아동교육연구단[*] 서울대 심리학과[**] 호서대 산업심리학과[***] 한국교육과정평가원[****]
광운대 산업심리학과[*****] 삼성인력관리위원회[*****] 성균관대 응용심리연구소[******]

본 연구에서는 영역참조적 검사제작의 논리를 적용하여 5급 국가고시(행정고시, 외무고시, 기술고시) 지원자들의 공직적격성을 측정할 수 있는 검사(PSAT: Public Service Aptitude Test)를 개발하였다. 그 동안 심리학자들이 제작한 검사는 거의 예외없이 규준에 비추어 검사를 해석하는 규준참조적 검사였다. 그에 반해서 본 연구에서는 이분모형에 따라 검사점수가 해석되는 영역참조적 검사를 제작하는 과정을 제시하고 있다. 이러한 검사에서 가장 중요한 것은 측정하고자 하는 영역의 정의와 명세이다. 제작 순서로서, 우선은 공직적격성의 영역을 파악하기 위한 이론적 분석과 심층면접을 통한 직무조사를 하였다. 결과로 언어능력, 상황판단능력, 자료해석능력, 사회상식 영역이 도출되었다. 문항 제작과정에서, 지원자 가운데 10%-15%가 합격하면서 그들이 전체 문항의 80%를 맞출 것이 요구되는 이분모형의 조건에 따라 문항의 곤란도가 설정되었다. 이러한 조건을 충족하는 검사를 개발하기 위하여 문항반응이론의 1모수모형을 사용하였다. 또한 제작된 검사의 영역참조적 신뢰도로서 두가지 방식을 사용하였다. 첫째는 이 검사에서 80% 이상 맞는 사람의 비율이 과연 10%-15%인지의 검토이고 둘째는 80%를 기준점으로 할 때 합격 불합격 판정이 두 개의 동형검사간에 일치하는 정도를 사용하였다. 타당화의 범위는 구성개념타당도로 제한하였다. 즉 영역간에 충분한 변별이 있는지 그리고 영역설계에 따라 검사의 1차원성이 성립하는 정도를 검토하였는데, 영역간 척도의 상관은 현저하게 작아서 변별성을 보여주었고 각 영역의 1차원성은 영역설계시 세부화의 정도에 따라 달랐다.

주요어 : 영역참조검사, 준거참조검사, 규준참조검사, 문항반응이론, 척도연결

* 연구자에게 연락은 smlyhl@chollian.net로 하기 바람.

그 동안 많은 심리검사들이 주로 규준참조(norm-referenced)방식으로 사용된 데 반해서 이 연구에서는 영역참조(domain-referenced)방식의 검사이론과 적용을 보이고자 한다. 즉, 영역참조적으로 사용될 검사의 제작논리를 적용한 사례를 제시하고 있다. 국가고시제도 개편의 일환으로 중앙인사위원회(2000a)에서 계획한 공직적격성 검사개발에 심리학자들이 참여하여 영역참조방식으로 사용될 검사를 제작하였다. 규준참조방식은 적절한 표본의 사람들에게서 구한 규준(norm)을 참조해서 검사를 해석하는 방식이고 영역참조방식은 목표로 하는 내용영역을 참조해서 검사를 해석하는 방식이다. 영역참조검사라는 용어의 원조로서 준거참조검사라는 용어가 '70년대에 발생하여 오랜동안 사용되어 왔으나 점차로 용어가 변해가는 과정에서 내용참조적 검사, 목표참조적 검사, 및 영역참조적 검사와 같은 명칭들이 사용되어 왔다. 최근에 Anastasi와 Urbina(1997)는 영역참조적 검사라는 용어가 더 정확하다는 이유로 이 용어를 권고하였고 이 글에서도 그것을 따른다.

규준참조검사 방식과 영역참조검사의 구분은 전자가 사람간 비교(상대평가라고도 함)에 중점을 두는 반면에 후자는 어떤 내용을 중심으로 평가(절대평가라고도 함)하느냐 하는 것이다. 따라서 규준참조방식에서는 검사받는 사람이 속한 모집단(population)이 어느 것이냐가 중요하고, 영역참조방식에서는 검사받는 내용의 전집(全集, domain)이 어느 것이냐가 중요하다. 일반적으로 모집단내에 있는 "사람들간 비교"하는 것을 상대평가, 내용전집에 비추어 개인을 "평가"하는 것을 절대평가라고 하지만 검사이론의 전문용어로는 규준참조검사, 영역참조검사가 된다. 여기서 영어의 "domain"을 "전집"이라고 번역하는 것이 절대평가의 분위기를 살리는데 적절하겠으나 현재 국내에서는 대체로 "영역"으로 번역하고 있다. 따라서 이 글에서는 영역과 전집을 상호교환적으로 쓰기로 한다.

영역참조방식의 검사가 심리학과 교육학 분야에서 단순히 개인차가 아니라 내용이 가지는 의미의 관점에서 개인을 평가하기 위해서, 그리고 내용영역에서 어떤 수준에의 숙달여부를 평가하기 위해서 이미 '80년대에 발달하였다 전자가 영역참조검사의 연속모형이라면 후자는 이분모형이다. 연속모형의 경우 어떤 측정대상의 구성개념에서 개인들의 수준이 다수준인 이해력, 비평적 사고력, 감식력, 및 독창력 등일 경우 적합하다. 즉 개인성장에서 한계를 둘 수 없는 영역에서의 평가에 적절하다. 그러나 어떤 기본적 수준을 필요로 하는 자격시험에서는 최소한 수준이 명시되고 그 수준을 넘느냐 미달하느냐 하는 것만이 관심의 대상이므로 연속모형은 현실에 맞지 않는 검사이다. 자격시험의 결과로 점수의 전 범위에 걸쳐 개인차가 드러나는 것은 사실이지만, 그러한 개인차를 극대화하는 목적으로 개발된 것이 아니고 어떤 기준이 되는 점수 즉, 분할점수 선상에서 이 쪽이냐 저 쪽이냐의 구분 가능성을 극대화하기 위해 개발되는 것이 이분모형에 의한 검사이다.

이 연구에서 개발하는 공직적격성 검사는 이름에서 암시되듯이 응시자 개인의 자질이 공직에 적절/부적절 여부를 보고자 하는 것이다. 공직적격성 검사는 본검사(2차시험)를 볼 수 있는 자격을 부여하기 위한 시험이기 때문에 공직에서 필요로 하는 영역을 바탕으로 하여 어떤 기준을 넘느냐 미달하느냐의 관점에서 개인을 평가하는 것이 목적이다. 따라서 규준참조방식의 사용이 아니고 영역참조방식, 그 중에서는 연속모형이 아닌 이분모형이 된다.

이 검사의 모형이 지원자를 "합격/불합격"으로 이분하는 이분모형이므로 모형에서의 요구조건이 충족되어야 한다. 그 조건은 5급고시(행정고시, 외무고시, 기술고시) 응시자의 10%~15%가 1차 시험에 합격할 것으로 보고 합격자는 80% 이상의 득

점을 해야 한다는(중앙인사위원회, 2000b) 것이다. 이분모형에서 보통은 분할점수(cut-off score)만 설정되지만 이 연구에서는 선발률(selection ratio)이 함께 주어진 독특한 영역참조검사가 되므로 규준참조방식에 익숙한 심리학자들에게는 생소한 영역이 된다. 이러한 조건을 충족하는 검사제작 과정 및 결과에 대해 다음과 같이 기술한다. 우선 영역참조방식의 개념을 규준참조방식과 대비하여 제시하고, 이 검사에서 측정할 영역(내용전집)을 설정하는 과정, 문항제작시의 기술적(technical) 방식, 그리고 제작결과 평가의 순서로 서술한다.

규준참조검사와 영역참조검사

국내에서 검사에 대한 연구가 주로 규준참조검사에 대해서 이루어져 왔지만 이제는 영역참조검사도 본격적으로 연구해야 할 때이다. 이미 심리학회 분과내의 몇 개 분과에서 실시하는 자격시험은 물론 표준직업사전에 제시된 560개가 넘는 공인자격의 시험들 모두가 규준참조가 아닌 영역참조의 검사방식이지만 이러한 방식이 갖는 이론적 특성에 대한 인식이 부족하고 그에 따른 기술적 관행이 확립되어 있지 않은 형편이다. 검사의 기술적 수준을 이야기하는 두 개념은 신뢰도와 타당도이다. 이 두 개념은 검사의 사용방식이 규준참조이냐 또는 영역참조이냐에 따라 다른 내용을 가지는 경우가 있으므로 주의를 필요로 한다. 이 글의 성격상 "검사의 사용"과 "측정"을 상호교환적으로 사용한다.

규준참조측정과 영역참조측정간에 기본적 차이는 분산(variance)에 있다(Popham & Husek, 1969). 규준참조일 경우 개인간 비교이므로 점수의 분산이 중요하다. 영역참조일 경우 개인간 분산은 관심의 대상이 아니다. 그래서 점수의 분산에 의존하여 정의되는 전통적인 신뢰도 개념이 영역참조검사에 적합치 않음이 인식되어야 한다. 영역참조검사는 두가지 종류로 구분되는데, 그 하나는 어떤 분할점수(cut-off score)를 기준으로 하여 그 영역을 숙달했음과 그렇지 않음을 나타내는 이분적인 결정의 경우이고, 또 하나는 '수용가능한 수행범위'의 점수들을 산출해 주는 경우이다(Popham & Husek, 1969). 전자의 경우는 응답자들을 두개의 범주에 분류하는 이분모형(binary model)이라고, 후자의 경우는 전통적인 의미에서 실시해 오는 '측정'과 같은 연속모형(continuous model)이라고 한다(Graham & Bergquist, 1975). 이 연구에서 개발된 공직 적격성 검사는 이분모형에 따른 영역참조검사이다. 이분모형에서는 분할점수를 선정하는 것이 핵심이 되며, 검사의 사용은 곧 개인이 그 분할점수의 수준에 도달했느냐 아니냐의 판단을 내리는 것이다. 이 연구에서는 분할점수가 미리 주어진 대신 그 점수 이상을 맞는 사람의 비율이 일정한 %일 것을 요구하고 있다. 그러면 신뢰도와 타당도의 개념, 그리고 문항개발시 문항분석의 방식이 규준참조방식과 영역참조방식간에 어떻게 다른지를 보기로 한다.

신뢰도

규준참조검사는 개인차를 보기 위한 검사이므로 진정한 개인차이를 반영해 주는데 있어서 개인점수 측정의 오차가 적을수록 바람직하다. 따라서 신뢰도의 정의가 진점수분산/(진점수분산+오차분산)으로 주어진다. 그러나 영역참조검사의 경우 어떤 전집(domain)에서의 점수추정 또는 응답자에 대한 분류결정(classification)이 주목적이므로 신뢰도의 정의를 '진점수분산/관찰점수분산'의 개념으로 쓸 것인가에 대해서는 의문을 제기하게 된다.

원래의 신뢰도 개념이 검사자체가 아닌 검사점

수의 신뢰도이므로 주어진 상황에서 얻어진 점수의 신뢰도를 논의할 때 상황에 따라 그 개념이 달라질 수 있다. 일반적으로 우리가 알고 있는 신뢰도는 규준참조방식에서의 신뢰도이다. 예로서, 검사-재검사 신뢰도, 동형검사 신뢰도, 내적일관성계수(스피어만브라운 반분계수, α계수)들이 그에 해당하며 모두가 "검사점수"의 일관성을 의미한다.

그러면 영역참조검사의 상황에서는 신뢰도가 규준참조검사에서와 어떻게 다르게 정의되는지 보기로 한다. 우선 이분모형에 대해서 이야기하자면, 두 개의 동형검사 또는 반복된 검사(재검사)를 실시했을 때 각각의 점수에 근거한 의사결정의 일관성이 있어야 한다는 것이 Hambleton과 Novick(1973)의 주장인데 이러한 주장이 미국의 검사표준서 6판(AERA, APA, NCME, 1999)에서 표준 2.15에 다음과 같이 반영되어 있다.

> "… 범주 분류의 결정을 내릴 때에는, 응답자들이 재검사 또는 동형검사[보통 시간적으로 떨어져서 실시]를 실시받은 후에도 같은 방향으로 분류될 수 있는 비율(%)의 추정치가 어느 정도인지 제공되어야 한다"(p.35).

즉 의사결정 사이의 일치도를 신뢰도로 보고자 하는 것이다. 이와 같은 것은 "검사점수"의 특성에 대한 것이 아니라 "의사결정"에 대한 것임이 분명하다(Traub & Rowley, 1980). 따라서 규준참조검사에서의 신뢰도 정의 방식과는 다르다.

연속모형은 전집점수의 연속선상에 개인을 위치시키므로 개인간 비교가 아니라 절대영역(전집)에의 비교가 된다. 이 때는 전통적인 신뢰도 개념을 사용할 수 있다(Ebel & Frisbie, 1991). 그러나 전통적인 신뢰도가 대상으로 하는 진점수(T)보다는 전집점수(universe score)를 대상으로 하므로 일반화가능도이론(generalizability)의 틀 속에서 산출되는 계수를 신뢰도로 제시한다. Cronbach, Gleser, Nanda, 및 Rajaratnam(1972)은 '진점수' 대신에 전집점수(universe score) 개념을 정의하고 신뢰도를 일반화시키는 대상을 하나가 아닌 여러가지 전집(universe)으로 볼 수 있다는 것을 개념화하였다. 여기서 상이한 전집의 예를 든다면 문항의 전집, 양식의 전집, 평가자 전집, 실시시간의 전집 등을 들 수 있다. 전통적인 신뢰도를 계산하는 대상은 하나(진점수) 밖에 없지만 일반화가능도이론에서는 전집이 하나 또는 여러개로 정의됨에 따라서 이 계수의 계산이 달라진다.

이 연구는 이분모형을 사용하므로 한 검사의 사용결과로 얻는 합격/불합격의 의사결정이 동형검사를 사용해서도 일치한다는 '분류의 일치도'가 신뢰도로 사용될 것이다. 또한 지원자의 10%~15%가 80% 이상을 맞아야 한다는 모형상의 요구에도 일치해야 신뢰할 수 있는 검사가 될 것이다.

타당도

Hambleton(1984)은 영역참조검사의 타당화 방식을 크게 목표내(intra-objective) 방식, 목표간(inter-objective) 방식, 준거타당도, 및 실험으로 나누고 있으나 심리검사 표준서에서 제시하고 있는 타당화의 세 방식(내용관련, 준거관련, 구성개념관련)으로 보는 것이 더 편리하다. 그럴 경우 목표내 방식은 내용타당도가 주를 이루고 구성개념타당도가 약간 정도 포함되어 있는 접근이다. 목표간 방식은 구성개념 타당도의 접근이고 실험 역시 구성개념타당도의 접근이다.

그런데 영역참조검사에서는 규준참조검사에 비해서 내용타당도가 크게 강조된다는 것이 차이점이다. 이것은 측정하고자 하는 목표영역 자체가 검사의 준거가 되기 때문이다. 이 때의 준거는 심

리학의 일부에서 바람직한 결과(예: 업무성과)를 준거로 하고 그 예측을 목적으로 하여 검사를 사용하는 경우와 다른 개념이다. 이 때의 목표영역에는 대체로 행동, 어떤 수준의 학습성취, 지식의 조직화 등이 포함된다. 이 연구의 경우 정해진 4개 영역 즉, 언어능력, 자료해석능력, 상황판단력, 및 기본소양(사회상식)의 영역에서 문항들이 표집되었는가에 대한 질적인 판단이 내용타당도가 된다. 내용타당도의 평가를 위해서 보통은 다음의 세가지 검토를 하게 된다(Hambleton, 1984).

- 충분한 수효의 판단자 및 측정전문가들에게 판단 의뢰
- 각 문항들이 검사목표에 일치하는지 검토
- 검사문항들의 기술적 적절성 검토

따라서 문항분석 절차에서 늘 내용타당도를 염두에 두고 검토하게 된다. 이 연구에서는 문항의 개발은 세부분야별로 석사·박사 과정생들이 하였고 6명의 전문가들이 문항들의 내용타당도를 검토하였다. 영역참조검사에서는 재고자 하는 영역 자체가 측정의 중심이 되므로 그 영역에서 표집된 [내용타당도 있는] 문항들이 사용된 것을 검토한 후에, 측정의 구조를 보기 위하여 구성개념타당도가 검토된다. 영역참조검사가 많이 쓰이는 분야가 운데, 교육장면에서는 교육영역 자체가 목적이 되므로 준거타당도는 크게 논의되지 않는다. 그러나 산업장면에서는 업무성과라는 또 다른 목적에 비추어 검사를 예측변수로 쓰고자 하므로 준거타당도 역시 중요하다. 이 연구에서는 내용타당도와 구성개념타당도를 중심으로 타당화가 되었으며 준거타당화는 후속연구의 몫으로 남겼다.

문항분석

규준참조검사의 문항분석에서는 난이도가 고르고 분산이 큰 문항이 바람직하지만, 영역참조검사에서는 모형에 따라 문항분석의 지침이 다르다.

난이도

규준참조검사에서는 문항점수의 분산이 큰 것이 개인차를 잘 구분해 주므로 분산을 극대화하는 것이 목표이다. 정답형 척도의 경우 정답률이 p일 때 문항점수의 분산은 p(1-p)이고 이 값은 p=.5에서 최대값이 되므로 평균적으로 중간 정도의 난이도를 가진 문항들이 권고된다(예: 정답률 0.3~0.7 사이의 문항들).

그러나 영역참조검사에서는 난이도의 의미가 모형에 따라 다르다. 이분모형이면 분할점수 근방의 사람들을 잘 변별해 주는 수준의 문항들이 필요하므로 모든 문항의 난이도는 그 수준에서 동일해야 한다. 그러나 연속모형의 경우 검사점수간 차이가 검사점수의 전범위(full range)에서 의미있으므로 다양한 난이도의 문항들이 필요하게 된다. 본 연구는 이분모형이므로 난이도가 동일한 수준의 문항들이 필요한 경우이다.

변별도

문항의 변별도는 측정되고 있는 속성의 수준에서 상하에 있는 사람들을 어느 한 문항이 얼마나 날카롭게 구분하느냐 하는 정도이다. 속성의 수준에서 높은 응답자가 그 문항에서 높은 점수를 받고, 낮은 응답자가 그 문항에서 낮은 점수를 받는 것이 바로 좋은 변별도이다. 일반적으로 규준참조검사에서는 변별도를 나타내기 위해 변별지수(discrimination index)와 문항-총점상관을 많이 사용하는데 전자는 영역참조검사의 이분모형에서도 원용할 수 있다. 즉 어떤 분할점수를 전후해서 있는 응답자들을 구분해 주는 민감도로서 "합격집단에서 그 문항에의 정답률"에서 "불합격집단에서 그 문항에의 정답률"을 뺀 값을 사용할 수 있다. 그러

나 문항총점상관은 문항의 양호도 평가를 위해 내부준거(검사총점)를 사용하므로 외부준거(내용영역)를 중심으로 하는 영역참조검사의 정의에 비추어 볼 때 중요성이 감소된다. 즉 문항-총점 상관이 큰 문항들만 사용한다면 검사내 문항들간 공변하는 정도가 높음을 의미한다. 만일에 영역이 넓게 정의되고 많은 문항이 표집된다면 영역내 문항들간의 공변성은 어느 정도 수용할 수 있다. 그러나 영역이 세부화되어 서로 구분되는 세부영역에서 소수의 문항을 제작하는 경우에 높은 공변성은 세부영역을 대표해야 하는 문항간 구분가능성을 저해한다. 이 연구에서는 영역내 세부화가 있었으므로 문항-총점 상관에 대해서는 관대한 기준을 사용하였다.

내용영역 설정

영역참조검사에서 영역 설정이 모든 작업의 시작이다. 영역참조검사에서 내용영역의 세부적 설정은 전문가들의 판단에 의존한다. 이 연구에서는 우선 문헌고찰을 통해서 설정한 후 현장의 사무관들(5급 행정고시, 기술고시 합격자들)을 면접하므로서 영역에 대한 질적 검토를 하였다.

공직 적격성의 하위영역

공직 적격성은 다양한 특성으로 구성되어 있으므로 이 공직 적격성의 "모든" 측면을 이 연구에서 개발하는 1차시험(PSAT)으로 측정하는 것은 가능하지 않다. 본 연구에서는 지필방식의 적성검사인 PSAT에서 측정하고자 하는 영역을 선정하는데 다음의 원칙을 사용하였다. 첫째, 고시에서 2차 시험이 전문적인 지식의 정도를 측정하고 있으므로 1차시험에서는 초급관리자로서 필요한 일반 적성과 공직관련 적성을 측정한다. 둘째, 현재 직무에 대한 지식이나 기술보다는 인지적인 문제해결능력, 환경 적응적인 상황판단력과 같은 초급 관리자로서 필요한 잠재력을 측정하려고 한다. 이러한 능력들은 기존의 국가고시 1차 시험에 포함되었던 헌법, 한국사, 및 기타 지식 과목에서는 측정될 수 없는 내용들이다.

위의 두가지 원칙을 준수하면서 공직 적격성 검사는 언어능력, 자료해석능력, 상황판단능력, 상식의 4가지 영역의 능력을 측정하기로 하였다. 이것은 일반정신능력이론(예: Gardner이론, Sternberg이론)과 직무분석의 결과에 근거한 것이었다. 이 영역 중 언어능력과 자료해석 능력은 전통적인 지능이론 및 최근에 Gardner(1983)의 지적능력에 대한 다면이론에서 제안된 7개의 영역 중에서 공직자와 유관한 것으로 판단되는 언어능력, 논리·수리능력의 개념을 따라서 구성되었다. Gardner의 지능 중 타인의 기분, 기질, 동기, 의도를 파악하고 구분하는 능력인 개인간(interpersonal) 지능은 현 한국사회에서 인성으로 오해를 유발할 소지가 있어서 본 검사에 포함되지 않았다.

또한 Sternberg(1985)의 실무적 지능(practical intelligence)에 근거하여 실무와 관련된 영역을 추출하였다. 실무적 지능은 학업지능처럼 일반적인 것이 아니고 영역중심의 지능(Sternberg, 1995)이라고 할 수 있으며, 영역에서 주어지는 상황에서의 사실들을 발견하면서 [지식을 획득하고] 자신의 장단기 목표에 맞게 적절히 반응하는 능력(Wagner & Sternberg, 1986)이다. "practical intelligence"를 조직장면에서는 실무적 지능으로 번역하기도 하지만 직무관련의 활동이 아닌 다른 활동이 있는 영역에서는 실용지능으로 번역되기도 한다. 현재 삼성그룹의 경우 조직에 맞는 "상황판단력"이라는 검사를 제작하여 실무적 지능을 측정하고 있는데, 입사후 훈련에서의 "발표력"에 대해서 기본정신능력 검사에 가까운 정도의 타당도를 보이고 있다(이창우,

이순묵, 김명언, 김명소, 유태용, 1998).

본 연구에서도 상황판단능력을 또 하나의 측정 영역으로 하고 있고 여기에서 실무적 지능을 측정하고자 하였다. 그런데 실무적 지능의 측정은 업무상의 기본이 되는 특정 상황을 피검자에게 제시하고 그 해결책을 찾아내게 하는 방식을 취하게 된다. 이때 정답은 객관적 정답이 아니라 경험적 정답(empirical key) 즉, 실제 업무에 종사하는 사람이 합리적으로 생각하는 해결책을 정답으로 삼는 방식을 택하게 되는데, 현실의 운영상 정답에 대한 이견이 발생한다면 새로운 제도의 정착초기에 부담으로 작용하므로 아직은 때가 이르다고 판단하였다. 따라서 상황판단영역은 경험적으로 정의되는 실무적 지능보다는 논리력, 추리력, 및 문제해결력으로 정의되었다. 따라서 이 정의에는 일반적성과 실무적성이 중복되어 있다고 볼 수 있다. 보다 구체적으로 상황판단력은 업무관련 능력으로서, 주어진 상황을 정확히 판단하고 해결책을 제시하는 능력, 주어진 상황을 개선하기 위해 방해가 되는 긴장이나 갈등의 요소를 해결하는 능력들을 포함한다. 끝으로 기본성취도 측면에서의 소양을 측정해야 한다는 관점에서 상식의 영역이 설정되었다. 상식은 사회의 규범이나 약속을 이해하고 따르는 행동과 사회 생활을 하는데 필요한 상식과 사회 정보, 매너 등을 의미한다. 이상에서 설정된 4개 영역은 현장에서의 직무조사를 통해서 검증과정을 가졌다.

직무조사

앞서 공직 적격성의 영역으로서 일반적성에 속하는 2개 영역(언어영역/자료해석영역), 업무 상황판단능력, 그리고 기본적 성취도에 해당하는 상식을 합하여 4개 영역을 제시한 바 있다. 언어능력이나 자료해석능력이 일반적인 학습능력이라면, 후자의 2개 영역은 실생활 또는 업무 관련의 학습능력에 관련된 영역이라고 볼 수 있다. 상황판단력이 보다 추상적이라면 상식은 보다 구체적이다. 따라서 직무 조사에서는 업무와 관련된 상황판단능력과 상식의 구체적인 하위영역을 발견하는 것을 목표로 하고 현직에서 2-3년된 고시출신 공직자 52명에게 면접을 실시하였다.

면접시에는 공직자로서 업무를 수행하는 데 필요한 능력에 대해서 반구조화된 면접을 실시하였고, 피면접자가 응답한 내용은 크게 언어능력, 논리력 및 판단력, 자료처리 및 수리 능력, 그리고 일반 지식 혹은 상식으로 수렴되었다. 구체적으로 각 영역에 대한 피면접자의 반응은 표1과 같다. 특히 상황 판단력은 논리력, 문제 해결력, 기획력, 그리고 의사결정력을 포함하고 있는 것으로 나타났다. 상식에서는 컴퓨터 관련 지식, 법률지식 그리고 국제적인 감각 등의 세부영역이 추출되었다. 그러나 대체적인 피면접자들의 반응은 특정한 일부 영역의 지식보다는 보다 광범위한 지식이 요구된다는 의견을 보였다.

피면접자들은 공직 적격성 검사의 철학과 현행 고시의 개편 방향에 대하여 전반적으로 긍정적인 의견을 보였다. 그리고 공직자로서 요구되는 능력도 앞서 개념적으로 설정된 영역인 언어력, 자료해석력, 상황판단력, 상식과 대체로 일치되었다. 이러한 영역은 일본이나 영국에서 시행되는 공직자 선발 검사 그리고 민간 기업인 삼성에서 실시하고 있는 SSAT의 영역과 유사하며 공직을 수행하기 위한 기본적인 능력으로 간주할 수 있다. 현직 공직자들의 효과적인 업무수행에 기본이 되는 능력이라고 생각하고 있는 영역이 이번에 연구에서 개발하고자 하는 공직 적격성 검사의 하위영역에 모두 포함되어 있다고 결론 내릴 수 있다.

표 1. 공직자로서 필요한 능력에 대한 사무관들의 반응

영 역	반 응
언어 능력	- 언어 이해력 (요점 판단)과 표현력 - 의사소통능력(타 영역 사업에 대한 이해 및 자기의 사업에 대한 설명력) - 자신이 알고 있는 지식을 종합/통합할 수 있는 능력
자료처리 및 수리 능력	- 자료분석 능력 : 통계 등 수치정보에서 정보 추출 - 정보분석 능력 - 수많이 제시되는 자료 중 필요한 자료를 추출하는 능력
논리력/ 문제풀이/ 판단력	- 논리력 : 귀납/연역 추리력 - 문제해결 능력 - 의사결정 능력 - 기획력 - 기존의 지식을 새로운 영역에 응용하는 능력
현대 사회에 필요한 지식과 법률 지식 (사회상식)	- 컴퓨터 관련 지식: 문서작성, 엑셀, 파워포인트 - Internet 지식, 인터넷 검색 - 법률지식, 법적용 mind : 행정법, 헌법 등 전체 법체계를 이해하고 있는 정도, 권리·의무관계, 법적 절차관계, 민법총칙 수준 정도, 질서 및 법 존중자세 - 국제감각 : 세계사/지리 등 상식수준 - 상식의 난이도 문제 : 시사적이며 2차 시험과 중복되지 않는 것
기타	- 창의력(아이디어 발상력) - 새로운 것에 대한 학습능력 - 종합적 사고력 - 분석력

문항분석의 논리

문항분석시의 조건

이 연구에서는 영역참조검사의 이분모형이 요구되고 있다. 즉, 어떤 능력범위 내에 있는 응시자들이 어떤 수준의 점수를 받도록 검사를 제작한다. 따라서 이미 분할점수가 주어진 것이며, 단순히 응시자 전체를 고르게 변별하고자 하는 것이 목적은 아니다. 검사의 제작에 있어서 응답자들을 전체 영역에서 고르게 변별하는 문항들의 제작보다는 제시된 조건에 충실한 문항들의 제작이 필요하다. 이 검사의 경우 "응시자 모집단에서 상위 10%~15%의 사람들의 마지막 사람이 4개 영역 평균점수가 80%정도가 되도록 한다"는 조건이 주어져 있다.

이러한 조건을 만족시키기 위해서는 두 가지 논리가 필요하다: 응시자 능력수준과 검사내용간 연결, 합격/불합격을 가르는 분할점수(cut-off score)를 문항에 연결.

응시자 능력수준과 검사내용간 연결

응시자의 능력모수의 어떤 수준이 검사 내 어떤 수준의 문항과 잘 일치하는 가를 찾아야 한다. 이것을 내용참조화(content-referencing)라고도 한다(Mislevy & Bock, 1986). 그러기 위해서는 응시자에 대한 측정체계(metric)와 문항에 대한 측정체계가 같아야 한다. 검사제작의 기본이론이라고 할 수 있는 고전검사이론에서는 문항의 정답률과 응시자의 점수간에 측정체계가 일치하지 않는다. 즉 정답률은 전체 응답자가운데 정답을 한 사람들의 비율이고, 응시자의 점수는 개인응답자가 답을 한 전체 문항들 중 정답을 한 문항의 수효(또는 비율)가 된다. 따라서 문항의 정답률과 응시자점수 간에 측정체계의 연결이 없다.

반면에 고급이론인 문항반응이론에서는 주어진 문항들에 비추어 개인응답자의 능력에 대한 모수를 직접 추정하며, 개별문항의 난이도(여기선 difficulty) 모수 역시 주어진 응답자 표본에 비추어 직접 추정한다. 5급 국가고시에 대한 전체응시자의 능력척도에서 일정한 위치에 있을 것으로 추정되는 집단에게 검사를 실시하고 이 집단의 능력척도(능력분포는 정규분포를 가정)에서 위치(location) 모수와 표준편차의 모수를 "일정한"값으로 부여하면 문항모수추정치와 능력모수추정치가 산출되고 응시자 능력수준과 검사내용간 연결이 된다.

합격자 집단의 최하점수를 문항에 연결

문항반응이론(Baker, 1985)에서는 능력 모수를 θ라하고 표준정규분포인 Z분포를 따르는 것으로 가정한다. 이 연구에서는 합격자를 응시자 가운데 상위 10%~15%라고 설정하고 있다. 응시자집단에서 상위 10%이면 능력 모수의 최하값은 Z분포에서 볼 때 θ_a=1.28이 된다. 그러나 합격자들을 충분히 잡아 상위 15%까지도 가능하다면, 이들에게서 능력 모수의 최하값은 Z분포에서 볼 때 θ_b=1.04가 된다. 그러면 능력이 θ_a나 θ_b의 수준에 있는 사람(합격자중 최하위 사람)이 "맞출 확률"이 80%인 문항들로 된 검사를 만들면 된다. 실제로는 수용상의 융통성이 있으므로 좌우로 10%씩 수용의 폭을 설정할 때 맞출 확률이 70%~90%가 되는 문항들이 필요하다. 이러한 조건은 문항반응이론의 1모수 모형을 사용하면 아래와 같이 표현된다.

$$p = \frac{1}{1+e^{-(\theta-b)}} \quad \cdots\cdots\cdots\cdots ①$$

따라서 전체 응시자의 능력척도에서 중앙에 있을 것으로 추정되는 집단의 능력모수에 대한 평균과 표준편차를 각각 0과 1로 할 때 바람직한 문항들의 곤란도는 -1.157과 0.433 사이에 있게 된다.

문항곤란도 b의 범위: -1.157 ~ .433 · · · ②

여기서 문항반응이론의 2모수 이상을 사용할 수 없는 이유는 다음과 같다. 식①에서 주어진 정보는 p와 θ밖에 없다. 만일에 다양한 변별도를 나타내는 변별도모수 a가 추가되면 a와 b중 하나를 일정한 값으로 고정해야 나머지 한 값이 결정된다. 즉, 정보부족으로 모수의 과소식별(underidentification)이 발생한다. 만일에 a=1로 고정하면 b가 결정되고 이것은 식①과 같은 1모수모형이 된다.

문항의 양호도 판단

우선 개발된 문항들의 양호도에 따라 1군(Group1)과 2군(Group2)으로 분류하기로 하였다. 전체 문항중 1군과 2군으로 분류된 문항을 중앙인사위원회에 제출하는 문항에 포함시켰다. 1군과 2군

을 분류하기 위한 기준은 다음과 같다.

1군 문항 선정기준

1군 문항 선정절차는 4단계로 나뉜다. 1단계에서는 극단적인 문항들을 제거하며 2단계에서는 남녀를 차별하는 문항을 제외하며, 3단계에서는 식 ②에서의 곤란도 추정치를 참고하여 1군 문항 후보를 선정한다. 4단계에서는 1모수모형에의 적합도 및 문항의 내용을 참고하여 최종적으로 1군으로 선정한다. 그러면 각 단계를 좀더 세부적으로 설명하기로 한다. 1단계와 2단계는 1차 실험평가(7월)와 2차 실험평가(10월) 간에 척도연결을 하기 전에 각 검사별로 분석하여 개략적으로 판단을 하는 단계이다.

1단계: PSAT는 극단적인 사람들을 선별하기 위한 것이 아니다. 따라서 극단적인 문항을 제거하기 위하여 정답률이 0.90 이상 0.05 이하이거나, 문항점수와 총점간 점이연(point-biserial) 상관이 0.20이 안되거나, 각 자료세트(영역별로 1차-A, 1차-B, 2차-A, 2차-B)를 문항반응이론의 소프트웨어인 PC-BILOG(Mislevy & Bock, 1986)로 분석할 때 곤란도가 +3과 -3(표본의 능력모수 추정치의 평균과 표준편차를 0과 1로 했을때)을 초과할 때를 극단적 문항으로 보고 제외한다.

2단계: 남자에게 불리하거나 여자에게 불리한 문항은 즉, 집단에 따라 차별적으로 기능하는 문항은 PSAT의 공정성을 저하시키므로 제외한다. 그 판단을 위해서 집단이 복수일 때 사용하는 문항반응이론의 소프트웨어인 BILOG-MG(Zimowski, Muraki, Mislevy, & Bock, 1996)에서의 "DIF"기능을 사용하였다. 이 기능에 의하면 남녀집단 구분이 있다고 할 때의 χ^2와 없다고 할 때의 χ^2간에 후자가 유의하게(χ^2차이 검증) 크면 구분이 있는 것이 모형적합도를 유의하게 증가시키므로 남녀차별 기능이 있는 것으로 보고 어느 성(性)에서 더 어려운 문항은 불리한 문항으로 간주한다.

문항차별기능은 이분문항의 경우(예: Lord, 1980)에서 다분문항의 경우(예: Collins, Raju, & Edwards, 2000)에 이르기까지 잘 연구되어 있다. 이분문항의 경우도 초기에는 문항 하나씩에 대한 차별기능을 검토하는 방식에서 지금은 문항집합 즉, 하나의 검사를 동시에 검토하는 방식까지 개발되었고 이 연구에서도 이 새로운 방법을 사용한 것이다. 이분문항의 경우 문항의 차별기능은 빈도표분석(contingency table analysis)에서 이분문항점수(1, 0)와 집단소속(남자, 여자)이라는 두 분류변수간에 상호작용이 있는가 없는가의 문제이다. 즉, 상호작용이 있으면 그 이분문항은 집단에 따라 쉽게 또는 어렵게 작용하므로 차별기능을 한다고 하는 것이다. 통계적 검증 역시 빈도표분석에서와 같이 χ^2검증을 사용한다.

1개 문항에 대한 검증에 비해 복수의 문항으로 갈수록 χ^2값도 커지고 자유도 역시 증가한다. 추정과정에서, 차별기능이 있다고 가정할 경우에는 입력자료에서 정보를 많이 사용해서 추정하므로 모형과 자료간의 합치도인 χ^2값이 작은 반면에, 차별기능이 없다는 제약을 가할 경우에는 앞서 차별기능을 허용할 경우에 비해서 모형이 간명하게(parsimonious)되므로 χ^2값이 증가한다. 이 두 χ^2값의 차이 역시 χ^2분포를 따르고 자유도는 "문항의 수효 x 모수의 수효"가 된다. 이 차이가 유의하지 않으면 문항의 집합 전체로서 차별기능이 없다는 모형이 유지되는 것이다. 만일에 두 모형간에 χ^2 차이가 유의하면 검사내에서 적어도 하나의 문항은 차별기능을 한다고 보아야 할 것이다. 즉 모수값이 집단간에 특별히 크게 차이나는 문항을 우선적으로 제거하고서 다시 같은 절차를 밟아 두 모

형을 비교해 본다. 이러한 계속적 절차를 통해서 두 모형간에 더 이상의 유의한 차이가 없을 때까지 차별기능문항을 골라내게 된다.

각 자료세트에서 위의 1단계와 2단계를 통과한 문항들만을 가지고 1차 실험평가의 문항들과 2차 실험평가의 문항들간에 척도연결을 하였다. 3단계 이후는 척도연결 후의 자료에 대한 판단이다.

3단계: 하나의 측정체계로 연결된 문항들에서 각 문항의 곤란도에 대한 점추정치(point estimate)를 중심으로 한 95% 신뢰구간을 설정한다. 즉, "점추정치±1.96(표준오차)"로부터 95% 신뢰수준의 구간추정치를 얻는다. 이 값이 바람직한 모수치의 범위인 [-1.157, 0.433]과 조금이라도 만나면 바람직한 곤란도 수준의 문항이 될 가능성이 매우 높으므로 1군 후보로 선정한다.

4단계: 1군 후보 문항들이 문항반응이론의 1모수모형에 잘 합치되어야 진정한 1군 문항이라고 할 수 있다. 합치도의 판단에서 영가설은 "1모수모형에 잘 맞는다"이다. 이것이 기각되면 1모수모형에 합치하지 않음으로 본다. 그런데 이 기각 판단은 보수적으로 하는 것이 이 연구의 목적에 적절하다. 따라서 영가설(1모수모형에서 잘 맞는다)이 기각되기 위해서는 확률치가 0에 가까울 때를 기준으로 하기로 하였다. 그렇지 않은 경우 모형 합치도에 대해서는 크게 우려하지 않기로 하였다.

2군 문항 선정기준

1군으로 선정되지 않은 문항들 가운데 다음의 기준들 중 하나만 만족시켜도 2군으로 선정한다.

- 곤란도의 구간추정치가 바람직한 모수치의 범위인 [-1.157, 0.433]에서 아래위로 약간만 이탈한 문항.
- 1군 후보이긴 했으나 1모수모형에의 합치도가 나빠서 최종적으로 1군이 되지 못한 문항.
- 1군 선정을 검토할 때 1단계에서 탈락되긴 했으나 그 때 BILOG로 계산한 곤란도 추정치가 -1과 +1 사이에 있는 문항. 이런 문항들은 적어도 극단적인 문항들은 아니다.
- 1군 선정을 위한 1단계 검토시에 점이연상관이 0.2 이상이었던 문항. 즉 어느 정도 영역내 문항들과 공변하는 문항들을 말한다.

척도연결

척도구성

척도의 구성은 검사개발의 일정상 1차 실험평가(7월)와 2차 실험평가(10월)로 구분된다. 이 때 영역별 척도내 문항수효는 표 2와 같다.

표 2. 척도구성

	1차	2차
영어	30	6(연결문항)+24(새문항)
상황판단	30	6(연결문항)+24(새문항)
자료해석	30	6(연결문항)+24(새문항)
상식	30	10(연결문항)+40(새문항)

척도연결의 논리

1, 2차에 걸친 실험평가를 완료한 후 문항의 척도(이때의 척도는 '측정체계'의 의미임)를 연결해서 최종적으로 1군과 2군을 선정하여 제출용 문항으로 하였다. 이 때 1차에서도 A형, B형이 있고 2차에서도 마찬가지였으므로 문항들을 하나의 측정체

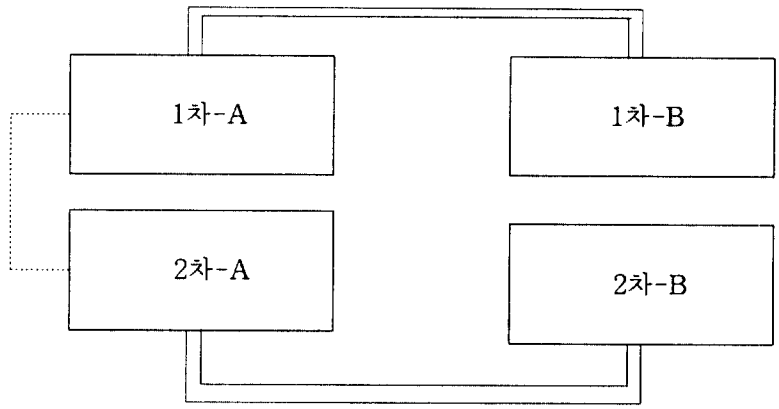

그림 1. 척도연결방식

계(metric) 위에 놓고서 모수추정을 하려면 척도연결이 필요하다. 척도연결(Keeves, 1998; Mislevy, 1992)은 크게 동등화(equating), 조정(calibration), 통계적 조절(statistical moderation), 예측(prediction), 그리고 사회적 조절(social moderation)로 나뉜다. 이 연구에서는 동등화를 하게 되는데 각 실험평가를 위해 개발된 문항들은 임의로 A형과 B형으로 나누었고, 한 장소내의 응답자들을 앉은 상태에서 홀수/짝수로 하여 홀수에 A형, 짝수에 B형을 실시했다. 이것은 동질집단간 동등화가 되므로 A형과 B형간에 동등화를 위한 연결(anchor, bridge)문항이 없어도 가능하다. 그러나 1차와 2차 사이에는, 문항의 곤란도는 물론 집단간 약간의 차이도 있을 수 있으므로 1차 평가에서 사용된 척노의 문항중 20%를 2차 실험평가의 척도에 포함시켜서 비동질 집단간 동등화를 시도하였다.

구체적으로는 척도연결용 집단("가교집단"으로 부르기로 함)의 능력모수치의 평균을 0, 표준편차를 1로 하고 4개 자료 즉, 1차-A형, 1차-B형, 2차-A형, 2차-B형 간에 문항모수의 척도연결은 그림1과 같은 방식으로 실시된다.

그림 1에서 ⌐¬ 는 동질집단 연결을 의미하며, 점선으로 된 ┊ 는 이질집단 연결을 의미한다. 즉, "1차-A"집단과 "1차-B"집단간 연결이 가능하고, "2차-A"집단과 "2차-B"집단간 연결이 가능하다. 그러면 1차의 한 집단과 2차의 한 집단을 연결하면 전체적인 연결이 가능하다.

"2차-A"의 척도에서 앞부분을 "1차-A"의 문항 중 대체로 20% 정도 뽑은 것으로 대체하여 척도연결용 검사를 만들어서 가교집단에 실시하였다. 척도연결용 검사의 구성은 상식을 제외한 세 영역에서 "1차-A"의 문항이 6개, "2차-A"의 문항이 24개였다. 상식에서는 "1차-A"의 문항이 10개 "2차-A"의 문항이 40개였다. 척도연결용 검사는 2차평가하는 시기에 실시되며 이 때는 새로운 문항들의 평가를 중심으로 하므로, 2차 문항들이 주로 포함되어 있고 1차 문항들은 연결문항으로서만 포함되었다.

가교집단으로부터의 자료수집 완료 후 1군 문항선정의 1단계(극단문항 제외) 및 2단계(남녀차별문항)에서 탈락한 문항을 제외한 나머지 문항들에 대해서만 척도연결작업에 들어갔다.

척도연결 및 문항분석

E여대와 SK대학교 학생 600여명이 가교집단

(bridge group)으로서 척도연결용검사에 응답해 주었고 1인당 1개 내지 2개 영역에 응답하였다. 2개 영역에 응답한 경우는 언어와 상황판단을 함께 하거나, 자료해석과 상식을 함께 하였다. 이러한 묶음은 앞으로 고시에서 실제 사용될 때와 같은 방식의 묶음이기도 하다. 척도연결시에 가교집단의 능력모수 평균치를 0, 분산을 1로 하였으며 그것을 기준으로 문항들의 모수치 추정 및 각 응답자들에 대한 능력점수(θ)가 산출되었다.

다음은 언어영역 1차 검사의 A형에 대해서만 문항분석표를 표 3에 제시하면서 해설하기로 한다.

표 3에서 정답률은 응답자 가운데 정답을 한 사람의 비율이다. 점이연상관은 문항-총점상관이다. 곤란도는 이 검사(1차검사 A형) 자료만을 가지고 문항반응이론의 1모수모형으로(BILOG 사용) 분석했을 때 산출되는 곤란도이다. 1차선별은 앞서의 세가지 지표가 극단적일 때 일단 그 문항을 이후의 검토에서 제외하기 위한 절차이다. 즉, 정답률이 너무 높거나(0.9 이상), 너무 낮거나(.05 이하), 점이연상관이 너무 낮거나(0.2 미만), 곤란도가 너

표 3. 1차 검사 문항분석표(A형)

	정답률	점이연상관	곤란도	1차선별	DIF선별	척도연결후곤란도	SE	(-1.96)SE	(1.96)SE	곤란도선별	χ^2	*Prob.*	모형적합도	문항등급*
1	0.96	0.19	-6.027	X										
2	0.74	0.22	-2.057			-1.171	0.292	-1.743	-0.599	O	8.3	0.401		A
3	0.73	0.28	-1.947			-1.070	0.292	-1.642	-0.498	O	6.7	0.467		A
4	0.75	0.29	-2.171			-1.275	0.299	-1.861	-0.689	O	1.1	0.991		A
5	0.77	0.44	-2.407			-1.596	0.209	-2.006	-1.186		11.3	0.078		B
6	0.99	0.17	-8.710	X										
7	0.34	0.34	1.290			1.893	0.279	1.346	2.440		7.9	0.341		
8	0.70	0.20	-1.682			-0.828	0.278	-1.373	-0.283	O	9.8	0.202		A
9	0.93	0.26	-5.210	X										
10	0.64	0.23	-1.188			-0.376	0.267	-0.899	0.147	O	6.8	0.564		A
11	0.23	0.11	2.351	X										
12	0.61	0.31	-0.862			-0.077	0.270	-0.606	0.452	O	4.9	0.769		A
13	0.51	0.15	-0.062	X										B
14	0.91	-0.20	-4.613			-3.691	0.217	-4.116	-3.266		19.8	0.003	X	
15	0.86	0.37	-3.644	X										
16	0.94	0.37	-5.387			-4.667	0.390	-5.431	-3.903		12.2	0.002	X	
17	0.90	0.26	-4.363	X										
18	0.72	0.23	-1.893			-1.021	0.288	-1.585	-0.457	O	6.7	0.570		A
19	0.46	0.29	0.333			1.017	0.262	0.503	1.531		5.8	0.673		B
20	0.74	0.45	-2.114			-1.222	0.309	-1.828	-0.616	O	14.1	0.050		A
21	0.54	0.30	-0.325			0.415	0.264	-0.102	0.932	O	3.9	0.865		A
22	0.70	0.41	-1.682			-0.828	0.292	-1.400	-0.256	O	6.6	0.586		A
23	0.62	0.47	-0.954			0.259	0.187	-0.108	0.626	O	31.8	0.000	X	B
24	0.44	0.48	0.509			1.179	0.276	0.638	1.720		16.8	0.033		B
25	0.52	0.30	-0.193			0.536	0.263	0.021	1.051	O	9.3	0.231		A
26	0.75	0.41	-2.228			-1.327	0.314	-1.942	-0.712	O	11.9	0.064		A
27	0.25	0.25	2.176			2.700	0.301	2.110	3.290		9.3	0.228		
28	0.39	0.10	0.869	X										B
29	0.58	0.28	-0.636			0.130	0.264	-0.387	0.647	O	7.1	0.530		A
30	0.46	0.20	0.333			1.017	0.260	0.507	1.527		14.3	0.075		B

주. * 문항등급에서 A는 1군, B는 2군을 의미한다.
X: 기각 O: 선택, 지원자의 10%-15%가 70점-90점 받는 범위는 -1.157~.433임.

무 극단(+3 이상, -3 이하)일 경우를 참조하였다. DIF 선별은 소프트웨어 BILOG-MG에서의 기능을 사용하여 문항의 성차별기능을 검출하는 과정이다. 표 3에서는 차별기능하는 문항이 없다. 하한값과 상한값은 척도연결후 곤란도의 95% 신뢰구간이다. 이 값이 -1.157~.433에 떨어지면 적절한 곤란도의 문항으로 선별하였다. χ^2는 문항이 1모수모형에 잘 합치되는지의 측정치이다. 이 값에 대한 p값이 너무 작으면 그 문항에는 1모수모형이 잘 안맞는 것으로 보았다. 이상의 검토를 거쳐서 곤란도 선별 단계와 모형적합도를 통과하면 1군이고, 그 정도는 안되지만 1군에서 지나치게 멀지 않으므로 사용가능한 문항으로 판단될 때 2군으로 하였다.

개발된 전체 108개 문항중 1군과 2군으로 구성되는 사용가능한 문항은 상식영역(42개)을 제외한 나머지 3개 영역에서 각각 60개였다. 이러한 문항 분석 결과로 선택된 60개 문항들을 이 후에 "개발된 검사"로 언급하는데 문항수효는 표 4와 같다.

표 4. 개발된 검사의 문항들

	1차-A	1차-B	2차-A	2차-B	합계
언어	20	17	13	10	60
상황판단	19	15	12	14	60
자료해석	20	16	12	12	60
상식	13	11	9	9	42

개발된 검사의 신뢰도와 타당도

우선 신뢰도와 타당도를 보기전에 각 영역별로 4개 자료 세트(1차-A, 1차-B, 2차-A, 2차-B)내에 들어 있는 선택된 문항들(1군, 2군)에서 개인이 정답을 한 문항수를 개인점수로 하여 요약된 자료를 표 5에 제시한다. 각 자료 세트 내에 문항의 수효가 그 세트내 만점의 값이 된다.

신뢰도

제출되는 1세트와 2세트에 들어있는 문항들은 모든 1차-A, 1차-B, 2차-A, 2차-B의 자료세트에 분산되어 들어있다. 따라서 각 자료세트 내에서 제출용 문항들만을 가지고 신뢰도 분석을 하기로 한

표 5. 영역별 점수 요약

	1차-A				1차-B			
	N	문항수	평균	표준편차	N	문항수	평균	표준편차
언어	183	20	12.37	2.96	161	17	10.67	2.73
상황판단	187	19	9.91	2.97	168	15	7.14	2.56
자료해석	181	20	9.68	3.02	157	16	7.80	2.73
상식	194	13	7.97	1.82	175	11	7.05	1.98
	2차-A				2차-B			
		문항수	평균	표준편차	N	문항수	평균	표준편차
언어	221	13	7.35	1.79	227	10	6.47	1.84
상황판단	216	12	6.44	1.96	236	14	8.67	2.69
자료해석	218	12	6.47	2.62	218	12	6.78	2.69
상식	211	9	5.68	2.00	216	9	5.32	1.96

표 6. 80% 이상 맞은 사람의 비율(%)

	1차-A	1차-B	2차-A	2차-B	영역별 중앙값	전체 중앙값*
언어	11.48 (N=183)	18.01 (N=161)	10.86 (N=221)	13.22 (N=227)	12.35	10.89
상황판단	6.42 (N=187)	4.76 (N=168)	11.11 (N=216)	26.70 (N=236)	8.77	
자료해석	3.87 (N=181)	8.28 (N=157)	10.55 (N=218)	16.97 (N=218)	9.42	
상식	22.16 (N=194)	24.57 (N=175)	37.91 (N=211)	33.33 (N=216)	28.95	

주. * 전체 중앙값은 영역별 중앙값들 가운데 중앙값이므로 엄밀한 의미의 중앙값은 아님.

다. PSAT는 기본적으로 영역참조적으로 사용되므로 요구조건의 만족도 및 분류의 일치도가 신뢰도로 사용된다.

PSAT에서는 우선 지원자의 10%~15%가 80%이상의 점수를 받을 때 검사에 대한 신뢰도가 부여된다. 우리는 각 영역별로 4개 자료 세트(1차-A, 1차-B, 2차-A, 2차-B)에서 세트내 총점 기준으로 80%이상 맞추는 사람의 비율을 표 6과 같이 구하였다.

표 6에서 보면 80% 이상 맞추는 사람들의 비율이 언어 영역에서 10%~15% 이내에 있고, 상황판단 영역과 자료해석에서는 10% 미만이 되며, 상식 영역에서는 15%를 넘는다. 즉, 상황판단영역과 자료해석의 문항들이 상대적으로 좀 어려운 편이고 상식 영역이 좀 쉬운 편이다. 그런데 이 자료는 단지 실험평가에 지원한 사람들을 대상으로 한 것이며 모든 지원자가 4개 영역을 모두 다 응답한 것도 아니므로 4개 영역의 총점을 가지고 80%이상이 몇 명이 나오는지는 계산할 수 없었다. 그러나 이 비율에 대해서 각 영역별 중앙값에 기초해서 전체적 중앙값을 구하면 10.89%로서 10%~15% 범위에서 벗어나지는 않는다.

표 7. 분류비율의 일치도

	언어	상황판단	자료해석	상식
N	59명	45명	52명	57명
비율	84%	60%	92%	81%

다음은 각 영역에서 응시자들이 80% 이상을 받은 사람의 집단과 그 미만을 받은 사람의 집단으로 분류할 때 개인에 대한 분류가 A형 검사와 B형 검사간에 일치하는 정도는 분류의 일치도로서 PSAT의 영역참조적 사용에서 또 하나의 신뢰도가 된다. 1차 검사 때에는 A형 검사와 B형 검사를 모두 응답한 사람들이 없었고, 2차 검사때는 영역별로 약 50여명씩이 있었으므로 이들의 자료에 기초해서 분류비율의 일치도를 구하면 표 7과 같다.

일반적으로 일치도의 값이 85%일 때 높다고(성태제, 1995) 하는 것을 참고로 하면 언어 영역과 자료해석 영역에서의 일치도는 높은 편이고, 상식 영역도 높은 편에 가깝다. 단지 상황판단 영역만은 높지 않다.

타당도

문항개발이라는 이 연구의 목적상 구성개념타당도를 중심으로 하였고 준거타당도는 후속연구의 몫이 될 것이다. 구성개념타당도는 상이한 개념간 변별성과 유사개념간 수렴성으로 대별된다. 영역참조검사에서의 구성개념타당도는 영역구성의 설계에 따라 검토하게 된다. 이 연구에서는 4개 영역은 엄연히 구분되는 영역으로서 설정하였다. 따라서 4개 영역점수간 상관은 충분히 변별된다고 할 정도로 작아야 4개 영역간 변별성을 보여줄 것이다. 각 영역내에서도 문항제작시에는 복수의 소영역으로 쪼개어 각 소영역을 대표하는 문항들을 제작하였다. 이 소영역들이 얼마나 독특한 개념이냐에 따라서 영역내 1차원성이 성립할 수도 있고 그렇지 않을 수도 있다. 즉 소영역들이 독특할수록 영역내 1차원성은 기각될 것이고 그렇지 않을수록 1차원성이 성립할 수가 있다. 참고로 4개 영역내 소영역들을 보면 아래와 같다.

언어영역: 지문의 배경은 19개 분야, 문제의 유형은 16개 종류.
상황판단: 소영역은 3개, 문제유형은 큰 특징이 없음.
자료해석: 지문배경은 6개 분야, 문제유형은 2개
상　　식: 지문배경은 정부 17개 부처와 관련있음, 문제유형은 큰 특징없음.

이 자료에서 볼 때 상황판단이나 자료해석은 검사결과 어느 정도 독특한 소영역이 떠오를 가능성이 있다. 즉 영역내 1차원성이 기각될 가능성이 좀 있다. 그러나 언어영역과 상식영역에서는 워낙 작게 나뉘어 있으므로 독특한 소영역보다는 전체로서 하나의 내용 즉 언어능력, 상식수준을 나타내게 될 가능성이 크다.

4개 영역 간의 변별성

4개 영역을 네시간에 걸쳐서 모두 응답한 사람만을 가지고 4개 영역간 상관을 구하면 영역간 별별성을 볼 수 있다. 4개 자료세트별로 상관을 구하고(표 8) 마지막으로 4개 자료세트에 걸친 평균값으로(표 9) 제시하기로 한다.

표 8. 4개 영역간 상관

1차-A(127명)				
	언어	상황판단	자료해석	상식
언어(20문항)	1.00			
상황판단(19문항)	.33*	1.00		
자료해석(20문항)	.26*	.09	1.00	
상식(13문항)	.26*	.09	.15	1.00
1차-B(103명)				
	언어	상황판단	자료해석	상식
언어(17문항)	1.00			
상황판단(15문항)	.17	1.00		
자료해석(16문항)	-.05	.21*	1.00	
상식(11문항)	.26*	.16	.02	1.00
2차-A(172명)				
	언어	상황판단	자료해석	상식
언어(13문항)	1.00			
상황판단(12문항)	.26*	1.00		
자료해석(12문항)	.25*	.29*	1.00	
상식(9문항)	.24*	.27*	.32*	1.00
2차-B(182명)				
	언어	상황판단	자료해석	상식
언어(10문항)	1.00			
상황판단(14문항)	.21*	1.00		
자료해석(12문항)	.27*	.29*	1.00	
상식(9문항)	.05	.13	.04	1.00

* : $p \leq .05$

표 9. 4개 영역간 상관의 평균

영역(문항수)	언어	상황판단	자료해석	상식
언어(60)	1.00			
상황판단(60)	.24	1.00		
자료해석(60)	.17	.18	1.00	
상식(42)	.21	.16	.14	1.00

표 8을 보면 대체로 언어능력은 상황판단력, 자료해석력, 및 상식과 유의한 그러나 크지는 않은 상관이 있다. 다음으로는 상황판단력과 자료해석력이 유의한 상관이 있거나 없거나 한다. 표 8의 4개 행렬을 평균하면 표 9와 같다.

이 표에서 보면 언어능력은 상황판단 및 상식과 0.20을 넘는 크기의 상관이 있고 자료해석은 언어능력과 0.17, 상황판단과 0.18, 상식과 0.16의 상관이 있다. 이들 크기는 이 4개 영역간 관련성을 보여주긴 하지만 관계의 정도가 크다고 할 수는 없음을 의미한다. 따라서 4개 영역간에는 충분한 변별성이 있다.

1차원으로의 수렴성

우선 1세트와 2세트의 문항들중 1차-A, 1차-B, 2차-A, 2차-B의 각각에 들어있는 문항들만으로 1차원성을 검토하였다. 소프트웨어인 TESTFACT (Wilson, Wood, & Gibbons, 1991))에 의해서 1차원성을 검토한 결과는 아래와 같다. TESTFACT에서는 문항반응이론을 기초로 차원성을 검증할 수 있다. 그 검증의 논리는 완전정보 방식에 의한 최대우도법으로 요인분석을 실시하여 χ^2검증을 하는데 영가설은 "1차원모형과 2차원모형간에 합치도의 차이가 없다"는 것이다. 즉, 1차원모형에서 2차원모형을 택할 경우 합치도가 유의하게 향상되지 않으면 1차원모형이 유지되는 것이고 유의한 향상이 되면 1차원성이 기각되는 것이다.

표 10을 보면 예측된 바와 같이 언어와 상식영

표 10. 각 영역에서의 1차원성 검토

		문항수	응답자수	("1차원→2차원" 변화시) $\Delta\chi^2$	Δdf	p 값	1차원성
언어	1차-A	20	183	37.36	19	.01	기각
	1차-B	17	161	21.98	16	.14	유지
	2차-A	13	221	17.84	12	.12	유지
	2차-B	10	227	17.23	9	.05	기각
상황판단	1차-A	19	187	43.62	18	.00	대체로 기각
	1차-B	15	168	-	-	-	1차-B는
	2차-A	12	216	27.27	11	.00	자료행렬이 역이
	2차-B	14	236	35.93	13	.00	없음
자료해석	1차-A	20	181	42.23	19	.00	기각
	1차-B	16	157	20.91	15	.14	유지
	2차-A	12	218	30.93	11	.00	기각
	2차-B	12	218	26.54	11	.01	기각
상식	1차-A	13	194	14.98	12	.24	유지
	1차-B	11	175	20.70	10	.02	기각
	2차-A	9	211	22.14	8	.01	기각
	2차-B	9	216	10.44	8	.24	유지

역에서 1차원성이 반정도 유지되고 있고 상황판단과 자료해석영역에서는 1차원성이 대체로 기각되고 있다. 어떤 분명한 결론을 내리기는 어렵지만 영역설계에서 소영역으로 나눌 때의 독특성 여부가 잘 반영되는 것으로 보인다. 그러나 표 8과 표 9에서 본 바와 같이 4개 영역간 변별은 분명하므로 구성개념타당도의 핵심부분은 충족되고 있다.

요약 및 결론

본 연구는 국가고등고시의 1차 시험을 대치하는 공직적격성 검사 (PSAT)를 개발하는 데 목적을 두고 진행되었다. 이 검사는 국내의 심리검사에서 흔히 사용되지 않은 방식인 영역참조적 방식으로 사용되므로, 본 논문에서는 이러한 용노의 검사에 대한 논리를 제시하고 제작의 기법을 소개하였다. 이 연구에서는 규준참조적 검사와는 달리, 이분모형(합격/불합격)에 따라 검사점수가 해석되는 영역참조적 검사를 제작해야 한다는 것이었다. 또한 이 검사에서는 지원자 가운데 10%-15%가 합격하면서 그들이 전체 문항의 80%를 맞출 것이 요구하는 부가적인 제약이 존재하였다. 이러한 제약을 1모수를 사용한 문항반응이론을 도입함으로써 해결하였다. 즉 검사사용자가 원하는 능력에 해당하는 지원자들이 80% 정도 맞출 수 있는 곤란도의 파라메터를 가지는 문항을 선정하여 검사를 구성하였다.

문항의 선정은 극단적인 문항들과, 남녀에 차별적으로 기능하는 문항들을 우선적으로 제거한 후 나머지 문항들에 대해서 하나의 측정체계로 연결하여 검토하였다. 즉 각 문항의 곤란도의 구간추정치가 바람직한 곤란도 범위인 -1.157~0.433 사이를 지나게 되고 1모수모형에 대한 적합도가 나쁘지 않은 문항들이 1군 문항이 되고 나머지 문항들에 대해서 조건을 완화시켜 2군 문항들을 선정하였다.

영역참조적 검사를 사용함으로써 이 방식의 검사에 합치되는 두 가지 종류의 신뢰도(여기서는 reliability보다는 dependability의 개념) 방식이 제안되었다. 첫 번째 신뢰도는 검사에서 80% 이상 맞는 사람의 비율이 10%-15%인지의 여부로 결정하였고, 두 번째 신뢰도는 80%를 기준점으로 할 때 합격 불합격 판정이 두 개의 동형검사간에 일치하는 정도를 사용하였다. 제작된 검사에서 80% 이상 맞는 사람의 비율은 영역별로 볼 때, 언어영역은 12.35%, 상황판단영역은 8.77%, 자료해석영역은 9.42%, 그리고 상식영역은 28.95%로서 검사 전체로서의 중앙값을 대략적으로 구하면 10.89%였다. 이것은 10%~15%의 조건을 충족하는 것이다. 합격/불합격으로 분류에의 일치도는 언어영역 84%, 상황판단영역 60%, 자료해석영역 92%, 상식영역 81%로서 상황판단영역을 제외하고는 무난한 편이었다. 즉, 제작된 공직적격성 검사는 영역참조적 검사로서의 신뢰도가 완벽하다고 할 수는 없어도 비교적 높았다.

본 검사의 타당도는 검사의 목적에 맞도록 예측타당도를 산출해야 되지만 그러한 타당도의 계산을 위하여서는 검사 시행 후 상당한 시간이 경과되어야 하므로, 현 단계에서는 계산이 불가능하다. 따라서 타당화의 범위는 구성개념타당도로 제한하였고, 4개 영역간에 충분한 변별이 있는지 그리고 영역설계에 따라 영역내 1차원성이 어느 정도 성립하는지를 검토하였다. 영역간 척도의 상관은 현저하게 작아서 변별성을 보여주었다. 1차원성은 검사발주자의 요구조건에 맞는 문항의 곤란도를 결정하기 위한 틀로서 사용된 문항반응이론에서의 중요한 가정이다. 그럼에도 실제로 각 영역내 설계시에 소영역들이 뚜렷하게 부각되는 경우에는 1차원성이 기각되었다. 특히 상황판단영역

과 자료해석영역에서 그러하였다. 이런 경우에는 각 소영역내에서 1차원성을 본다면 성립할 가능성이 높을 것이다. 그럴 경우 그 두 영역에서는 영역별로 요인분석을 하여 차원의 수효를 결정하고 각 차원내에서 다시 앞서 사용된 소프트웨어 TESTFACT를 사용한다면 1차원성이 지지될 것이다. 그러나 그렇게 하기 위해서는 문항선정시에도 그러한 소영역별로 문항을 균형있게 배분하는 노력이 선행했어야 한다. 그러나 이 연구에서는 영역단위로 문항을 선정하였으므로 이 점은 이 연구의 한계라고 해야 할 것이다.

또한 본 검사가 기존의 행정고시 1차 점수와는 낮은 상관(.00~.19)을 보였고, 오직 상식 영역 만이 비교적 중간크기에 가까운 상관(.10~.32)을 보였다는 것도 적성을 중심으로 한 본 검사가 지식을 주로 측정하는 현행 고시와는 차별적임을 보여주고 있다.

고시응시 후보대학생, 고시 합격자(수습사무관), 재직자의 세 집단이 본 검사의 점수에서 차이가 난다는 사실도 간접적으로는 공시(concurrent)타당도로서 제시될 수 있는데, 언어력을 제외한 영역에서 재직자가 가장 높은 점수를 보여 주었다(표 11 참조). 이는 본 검사가 대체로 공직에의 적격성을 측정하고 있음을 시사하고 있다.

본 연구의 의의는 심리측정이론에 근거하여 공적 영역에서의 시험을 개발하는 과정을 명세하여 주었다는 데 있다. 기존의 국가고시가 시행되는 과정에서 검사이론이라는 측면이 고려되지 않았기 때문에 기존의 검사들에 대한 심리측정적 평가를 할 수 있는 자료가 미비하다. 이러한 점은 검사에 대한 논리와 타당화에 대한 평가를 가능하게 하는 자료를 제공하는 것을 어렵게 함으로써 검사에 대한 객관성과 공정성에 대한 의문을 제기할 수 있게 한다. 본 연구에서 개발된 검사가 얼마나 양호한 검사인지는 추후의 타당화 연구에서 밝혀져야 하겠지만 적어도 검사의 개발절차와 타당화 절차에 대하여 검사이론에 충실한 절차를 따름으로써, 이후의 연구자들이나 테스트 사용자가 검사를 평가하고 개선하는 데 도움을 줄 수 있을 것이다.

방법론상에서 영역참조적 검사의 개발 절차에 대한 하나의 모형을 제시하였다는 데에서 본 연구의 또 다른 의의를 찾을 수 있을 것이다. 영역참조적 검사는 입학시험이나 채용시험과 같은 맥락에서 흔히 사용되어지는 검사이지만 그에 걸맞는 인식 속에서 검사가 개발된 경우는 많지 않다. 특히 본 연구에서는 내용영역 중심의 일반적인 조건 외에 부가적으로 합격률과 합격점수에 대한 기준이 제공되는 조건 하에서 문항을 선정하는 한 방법을 제공하였다.

본 연구에서 제안되었던 방식에 대한 평가는 추후의 연구에 의해 결정될 것이다. 특히 영역참조 검사 신뢰도의 경우는 더 많은 검사개발자에 의해 연구되어야 할 것으로 보인다. 그러나 영역참조적 검사를 개발하는 과정에서 이러한 유형의 검사를 개발하는 논리와 방법론에 대한 연구가 비교적 적게 이루어 졌다는 점을 고려하면, 본 연구는 이러한 연구를 촉발시키는 계기가 될 수 있을 것이다.

표 11. 척도연결후 세 집단의 평균점수

영역	대학생	재직자	수습사무관
언어	.15 (507)	.32 (120)	.38 (370)
상황판단	.00 (482)	.01 (127)	-.24 (378)
자료해석	.14 (470)	.35 (115)	.09 (370)
상식	.13 (478)	.72 (132)	.50 (381)

주. ( ) 안은 사람수

참 고 문 헌

성태제 (1995). 타당도와 신뢰도. 서울: 양서원

이창우, 이순묵, 김명언, 김명소, 유태용 (1998). 삼성직무적성검사 개발 최종보고서. 서울: 성균관대학교 산업 및 조직심리 연구소.

중앙인사위원회 (2000a). 2000년도 업무계획. 서울: 저자.

중앙인사위원회 (2000b). 공직 적격성 테스트 문제 개발 용역제안 요청서. 서울: 저자.

Anastasi, A. & Urbina, S. (1997). *Psychological Testing*. 7th Ed. Upper Saddle River, NJ: Prentice Hall.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, D.C.: Authors.

Baker F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology, 85*. 451-461.

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.

Ebel, R. L. & Fresbie, D. A. (1991). *Essentials of educational measurement*. (5th ed.). NJ: Prentice Hall.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligence*. New York, NY: Basic Books.

Hambleton, R. K. (1984). Validating the test scores. In R. A. Berk (Ed.) *A guide to criterion-referenced test construction*. Baltimore: The Johns Hopkins University Press.

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.

Keeves, J. P. (1998). Scaling achievement test scores. In Walberg & Haertel (Eds.), *The international encyclopedia of educational education*, (pp. 308-318). Pergamon Press.

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

Mislevy, R. J. (1992). *Linking educational assessment: Concept, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.

Mislevy, R. J. & Bock, R. D. (1986). *PC-BILOG: User's Guide*. Mooresville, IN: Scientific Software.

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

Sternberg, R. J. (1995). Theory and measurement of tacit knowledge as a part of practical intelligence. *Zeitschrift Für Psychology. 203,* 319-334.

Traub, R. E. & Rowley, G. L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement. 4,* 517-546.

Wagner, R. K. & Sternberg, R. J. (1986). Tacit knowledge and intelligence in the everyday world. In R. J. Sternberg and R. K. Wagner (Eds.), *Practical intelligence: nature and origins of competence in the everyday world*. NY: Cambridge University Press.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *Testfact: Test scoring, item statistics, and item factor analysis*. Chicago, IL: Scientific Software International Inc.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG*. Chicago, IL: Scientific Software.

한국심리학회지 : 일반
Korean Journal of Psychology
2001. Vol. 20, No. 2, 305-325

Rationale and Application of Domain-Referenced Test Development

Soonmook Lee[*] Cheongtag Kim[**] Myoung-So Kim[***] Hyunsoo Seol[****]
Tae-Yong Yoo[*****] Do-Hyoung Lee[******] Dae-Yeul Lim[*******]

Sungkyunkwan University[*] Seoul National University[**] Hoseo University[***]
Korea Institute of Curriculum & Evaluation[****] Kwangwoon University[*****]
Samsung Human Resources Development Center[******] Sungkyunkwan University[*******]

In the present study we applied the logic of domain-referenced testing to develop Korean Public Service Aptitude Test. Most of tests developed by psychologists can be categorized into norm-referenced tests. In contrast, we developed a domain-referenced test following the logic of discrete model having a cut-off score and selection ratio. The applicants are separated into two groups by the cut score(80% correct) and the proportion of passing applicants should be 10% ~ 15%. Applying the Rasch model of item response theory, we set an appropriate level of difficulty for items satisfying the two conditions for the test: 80% correct as the cut score, selection ratio 10% ~ 15%. The reliability of the test was defined by the conditions imposed by the test user. The validity of the test is limited to construct validity(convergent/divergent validity), leaving the examination of criterion validity for the follow-up study. Both reliability and validity were acceptable, although they were not outstanding.

Keywords : domain-referenced test, norm-referenced tests, item response theory, linking

1차원고접수 : 2001. 6. 11.
수정원고접수 : 2001. 11. 20.
최종게재결정 : 2001. 11. 26.