한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2010. Vol. 23, No. 2, 225-249

내재설계 평가센터의 신뢰도 및 타당도

허 창 구 신 강 현†

아주대학교

본 연구는 평가센터를 적용한 신입사원(1,249) 선발자료와 재직자(관리자 105, 일반사원 200) 평가센터자료를 이용하여, 평가센터의 신뢰도 및 타당도를 검증하기 위해 실시되었다. 특히 본 연구의 평가센터는 평가과제마다 특정 평가차원이 지정되어 있는 소위, 내재설계된(nested designed) 평가센터로서, 구조화 면접에서는 3명의 평가자가 3개의 평가차원(책임감, 적극성, 혁신성)을, 프레젠테이션에서는 2명의 평가자가 2개의 평가차원(정보분석력, 문제해결력)을, 그룹토의에서는 2명의 평가자가 2개의 평가차원(대인친화력, 의사전달력)을 평가하도록 설계되어 있었다. 이러한 내재설계 평가센터는 평가자에 의한 편향(rater bias)이나 평가과제 효과(exercise effect)가 우려되기 때문에, 평가의 신뢰도 및 타당도 검증을 통해 평가센터의 활용가능성을 확인할 필요가 있다. 일반화가능도 결정연구(D study)결과, 내재설계 평가센터는 양호한 신뢰도를 보여주었으며, 일반화가능도 일반화연구(G study)결과, 평가자의 편향은 거의 나타나지 않았고, 평가과제 효과가 매우 크게 나타났으나 평가차원의 변별성 또한 어느 정도 확인할 수 있었다. 확인적 요인분석을 통한 구성개념타당도 검증결과에서는 평가과제와 평가차원을 모두 고려한 통합모형이 가장 적합한 것으로 나타나 평가차원의 변별성을 보여주었다. 재직자의 평가센터 점수와 준거와의 상관관계분석 결과에서는 각 평가과제에 내재되어 있는 평가차원들이 어느 정도 차별적인 준거예측력을 보여주었으며, 전반적으로 구조화 면접에 속한 평가차원이 가장 포괄적인 준거관련성을 지녔고, 프레젠테이션에 속한 평가차원이 준거와 가장 관련성이 적었다. 논의에서는 본 연구의 의의와 제한점을 제시하였다.

주요어 : 평가센터, 내재설계, 일반화 가능도, D 연구, G 연구

† 교신저자 : 신강현, 아주대학교 심리학과, khs9933@ajou.ac.kr

취업경쟁이 심각해짐에 따라, 기업은 수많은 지원자 중 조직에 적합한 유능한 인재를 채용, 선발, 훈련, 유지할 수 있는 보다 정교한 도구를 필요로 하게 되었다. 더욱이, 업무 성과를 최대화하기 위해 조직이 팀제로 편성되고, 비즈니스가 점차 글로벌화 되면서 직무 간 경계가 점점 더 모호해짐에 따라(Cascio, 2002; Howard, 1995), 정해진 지식을 측정하는 것이 아닌 변화하는 직무에서의 성공 잠재력을 측정할 수 있는 도구를 필요로 하게 되었다. 한편, 지필검사(paper-pencil) 방식의 전통적 평가가 정보수집 능력, 문제 해결 능력, 창의력, 종합적 구성능력과 같은 고등 정신능력을 평가하기에 부적합하다는 비판 아래, 대안적 평가로서 수행평가(performance assessment)가 도입되어 활발하게 사용되고 있다(강애남, 이규민, 2006). 수행평가란 피평가자가 자신의 지식이나 기능을 이용해 실제 생활이나 인위적인 평가 상황에서 얼마나 잘 수행하는지 혹은 어떻게 수행할 것인지를 평가자가 서술, 관찰, 면접 등의 다양한 방법을 통하여 종합적이고 전문적으로 판단하는 평가방법이다(성태제, 2002).

최근 기업에서 종업원 선발, 진단, 개발을 목적으로 점차 그 사용이 확대되고 있는 평가센터는 다양한 모의활동(simulated exercises)을 통해 이루어지는 수행평가라고 할 수 있는데, 일반적으로 지원자가 지닌 다수의 역량(평가차원)을 다수의 평가과제를 통해 다수의 평가자가 평가하는 다각적인 평가(multiple assessments)가 이루어진다. 그러나 이러한 수행평가는 전통적인 지필검사와는 달리 채점 과정에서 피평가자의 능력 이외에도 평가자, 평가과제, 평가과제 수, 평가시기, 평가실시절차 등의 다양한 요소들 및 이 요소들 간의 상호작용에 영향을 받을 수 있다(이영식, 신상근, 2004). 특히, 피험자의 응답을 평가하기 위한 평가자의 개입을 반드시 필요로 하기 때문에 기존의 지필검사 중심의 전통적 평가에서는 고려되지 않던 평가자 국면(facet)을 측정에 포함하여야 한다. 그로 인해 수행평가 점수의 신뢰도가 크게 낮아지게 되고, 낮은 신뢰도는 수행평가 결과를 활용하는데 가장 큰 제한점으로 작용하고 있다.

한편, 애초에 독일에서 군 장교 선발을 위해 개발되었던 평가센터는 1940년대에 영국의 시민서비스 기관에서 비군사적 목적으로 처음 이용된 이후, 직업세계의 변화와 더불어 평가차원(dimension), 평가과제(exercise), 평가자의 유형 및 평가점수의 통합방식 등에서 다양한 변화를 겪어왔다. 평가센터의 활용목적에 있어서도 선발 목적 이외에, 종업원 진단을 위한 평가센터(Diagnostic Assessment Center)와 종업원 개발을 위한 평가센터(Developmental Assessment Center)로 발전되어왔으며, 평가센터의 구성에 있어서도 면접(interview), 검사(test), 설문(questionnaire), 모의활동(simulation exercise) 등 다양한 과제들과 다양한 평가차원이 조직의 상황과 필요에 맞게 다양한 형태로 구성되고 있다. 국내에서는 평가센터의 비용 및 시간적 문제로 인해 HPI(High Potential Individual), 임원 후보자 및 임원, CEO 후보자 등을 대상으로, 이들의 선발과 육성을 위해 사용되어왔으나, 최근에는 비용과 시간적 제약을 극복하기 위해, 좀 더 짧고 간단한 형태의 시뮬레이션을 개발하여 신입사원 선발에 활용하는 기업도 증가하는 추세이다(임대열, 2004). 이러한 신입사원을 대상으로 하는 평가센터는 대규모 지원자를 대상으로 하며, 지원자의 다양한 잠재역량을 평가해야 한다. 그로 인해, 복수의

표 1. 평가센터 설계의 구분

	교차설계		내재설계	
	역량1	역량2	역량1	역량2
평가과제1	●	●	●	
평가과제2	●	●		●

평가과제를 통해 평가차원들을 중복 평가하지 않고, 각 평가과제 마다 가장 잘 평가될 것으로 여겨지는 평가차원을 배정하여 평가하도록 설계하곤 하는데, 이러한 평가센터는 평가차원이 평가과제에 내재된(nested) 평가센터라고 할 수 있다[1]. 하지만, 평가차원이 특정 평가과제에만 국한되어 평가될 경우, 평가차원에 관계없이 평가과제의 특성에 따라 평가점수가 결정되는 평가과제 효과(exercise effect)의 영향으로 평가센터 연구의 오랜 숙제로 알려진 '취약한 구성개념 타당도'가 더욱 약해 질 수 있다.

더욱이, 이러한 수행평가 결과로부터 단지 피평가자의 능력을 관찰하고 다양한 정보를 얻는데 그치지 않고, 인사 · 선발과 같은 의사결정에 활용할 경우 수행평가 결과의 신뢰도와 타당도 검증은 더욱 중요해진다(남명호, 2002). 수행 평가의 활용도가 높아지면서 수행평가의 신뢰도를 추정하는 방식과 낮은 신뢰도를 제고시키기 위한 방안을 탐색하는 연구가 증가하고 있으나(지은림, 1999; 지은림, 김성숙, 2005), 수행평가 결과에 대한 측정학적 신뢰도와 타당도 검증에 관련된 연구와 논의는 여전히 미흡한 실정이다(강애남, 이규민, 2006). 따라서 수행평가를 기본으로 하는 평가센터에 대한 신뢰도 및 타당도 검증이 필요하다고 할 수 있으며, 특히 내재설계된 평가센터의 활용이 증가하고 있는 시점에서 이러한 유형의 평가센터가 지니는 신뢰도와 타당도를 확인하는 연구가 필요하다고 하겠다.

본 연구의 첫 번째 목적은 일반화가능도 이론을 이용해 수행평가 상황의 다양한 영향요소를 고려한 평가센터의 신뢰도를 확인하고, 일정한 신뢰도를 확보하기 위해 필요한 평가센터 설계를 제안하는 것이다. 평가센터와 같은 수행평가는 실제 행동에 대한 평가이기 때문에 평가의 타당도가 높은 반면, 다수의 평가자에 의한 평가이기 때문에 신뢰도를 확보하기 어려울 수 있으며, 표준화된 검사에 비해 시간과 비용 등이 많이 들기 때문에 실용도가 떨어지는 측면이 있다. 이렇듯 신뢰도와 실용도가 낮은 평가 방법은 현장에서 실행하기에 약점이 있기 때문에 믿고 활용할 수 있는 평가방법이 되기 어렵다. 따라서 평가센터와 같은 수행평가가 현장에서 올바르게 정착되기 위해서는 평가의 신뢰도와 실용도 문제가 해결되어야 한다(조재윤, 2009). 평가센터와 같은 수행평가 점수에는 평가자(rater), 평가과제(exercise), 평가차원(dimension) 및 이들의 상호작용과 같은 다양한 요인이 영향을 미친다. 따라서 수행평가의 신뢰도는 위와 같은 요소들의 영향을 고려하여 산출할 필요가 있다. 이에 Cronbach, Rajaratnam 그리고 Gleser(1963)는 다양한 요인이 영향 미치는 수행평가의 신뢰

1) 이러한 형태의 평가센터를 위한 별도의 명칭이 없는 관계로, 본 연구에서는 일반화가능도 연구 설계의 용어를 빌려 '내재설계 평가센터'로 칭하겠다.

도를 산출하기 위한 일반화가능도 이론을 고안했다. 일반화가능도 이론을 이용하면, 평가센터의 신뢰도를 확인 할 수 있다. 또한 각 요인의 영향력을 비교함으로써 신뢰도 확보를 위해, 이들 구성요소들을 몇 명(개)씩 포함시켜야 할지와 같은 평가센터 설계에 도움이 되는 정보를 확인함으로써 평가센터의 실용성을 향상시킬 수 있다.

본 연구의 두 번째 목적은 내재설계 평가센터의 평가차원 변별성을 확인하여 구성개념 타당도를 검증하는 것이며, 준거와의 관련성을 통해 내재설계 평가센터의 준거관련 타당도를 확인하는 것이다. 전술한 바와 같이 평가센터의 활용이 대규모 일반사원 선발에 까지 확대됨에 따라, 운영상의 문제로 몇 개의 모의과제를 수행하도록 제시되기도 한다. 이 경우 한 평가과제에서 여러 평가차원들이 평가되고, 각 평가차원들은 특정 평가과제에서 한 번씩만 평가되기 때문에, 평가센터 점수에 미치는 평가과제의 영향이 평가차원의 영향에 비해 커지게 된다. 기존의 평가센터 연구들에 따르면 내용타당도와 준거관련 타당도가 안정적으로 검증되는데 반해, 구성개념 타당도의 증거는 비일관적이다. 따라서 이러한 내재설계 평가센터 상황에서도 평가센터의 구성개념 타당도가 확보될 수 있는지 다시 말해, 평가차원들이 개별적으로 평가될 수 있는지를 검증할 필요가 있다. 만약 내재설계 평가센터에서 평가차원이 개별적으로 평가되지 않는다면, 평가센터는 더 이상 평가차원을 기반으로 설계될 필요성이 없다고 할 수 있다. 이는 평가센터 구축을 위한 첫 단계로서의 직무분석이나 역량모델링의 필요성에도 위협이 되는 결과이며, 평가센터 결과의 활용에 있어서도, 평가센터 전체 점수(Overall Assessment Rating, OAR)를 이용하는 인사선발 목적에 평가센터의 활용을 국한시켜야 하며, 각 평가차원 별 점수를 이용하는 개발(development)이나 진단(diagnostic) 목적으로는 활용을 자제해야할 것이다.

평가센터의 정의 및 형태

Thornton과 Rupp(2003)은 평가센터를 “최소한 하나 이상의 모의상황에서 여러 평가방식을 통해 보여준 지원자의 수행을 평가하는 방법”이라고 정의했다(p319). 평가센터는 50여 년 전 소개된 이후 다양한 평가요소들을 독특한 조합으로 구성한 선발과정을 통칭해왔다. 이러한 평가센터의 운영은 해당 기업이나 조직의 원칙(disciplines)과 상황(settings)에 따라 다르게 진행되었으며, 1975년에 가이드라인이 만들어지기까지 통일된 모습을 지니지 못했다. 1975년에 캐나다 퀘벡에서 개최된 평가센터방식 국제회의(International Congress on the Assessment Center Method)에서 평가센터 운영방식의 가이드라인을 개발하기 위한 국제 태스크포스가 구성되었는데, 이것이 익히 알려진 ‘평가센터 운영을 위한 가이드라인과 윤리적 고려사항(Guidelines and Ethical Considerations for Assessment Center Operations)’이다. 이 가이드라인은 실제 가장 잘 수행된 평가센터를 통합하여 만들어 졌으며, 이후 세 차례 수정을 거쳐, 2009년에 최근판(International Task Force on Assessment Center Guidelines, 2009)이 발표되었는데, 이 가이드라인에는 평가센터의 필수적인 요소가 표 2와 같이 기술되어있다.

본 연구에서 사용된 평가센터는 평가차원이 평가과제에 속해있는 내재설계이지만 평가센터 개발과정과 실시과정에서 아래의 모든 특

표 2. 평가센터의 기본 요소

1.	Job analysis/competency modeling 직무의 성공과 관련된 역량을 규명하기 위한 직무분석이나 역량모델링이 수행되어야 함
2.	Behavioral classification 지원자가 보여주는 행동이 의미 있는 범주(특성, 적성, 역량, 기술, 지식 등)로 분류되어야 함
3.	Assessment techniques 평가기법은 직무분석으로 결정된 역량을 평가하기 위한 정보를 얻을 수 있도록 설계되어야 함
4.	Multiple assessments 복수(multiple)의 평가기법이 사용되어야 함
5.	Simulations 평가기법에 직무관련 모의상황이 포함되어 있어야 함
6.	Assessors 복수의 평가자가 각 지원자를 관찰/평가해야 함
7.	Assessor training 평가자는 역량, 관찰/기록/분류/평가, 과제의 내용, 평정오류에 대한 훈련을 받아야 함
8.	Recording behavior and scoring 평가자에 의해 지원자의 구체적인 행동이 관찰과 동시에 기록되어야 함
9.	Data integration 평가자들이 정보를 통합하거나 통계적인 통합과정을 거쳐 지원자의 행동이 통합됨

성을 갖추고 있다. 따라서 내재설계 평가센터 또한 평가센터의 또 하나의 유형으로 볼 수 있다.

평가센터의 타당도

평가센터 방식은 지원자들과 직원들을 평가하고 개발하는 포괄적이고 탄력적인 도구로서, 복잡한 특성의 측정이 가능하고, 참가자들에 의해서도 공정하다고 인식되고 있으며, 편파효과(adverse impact)가 거의 없고, 다양한 준거(수행, 잠재력, 훈련성공, 승진)를 예측해주는 것으로 연구되었다(Gaugler, Rosenthal, Thornton, & Bentson, 1987). 따라서 현재 평가센터(Assessment Center)는 GE, MicroSoft, GM, IBM, Sony, Coca Cola, P&G, Kodak, Sears, Unilever 등 해외 선진기업에서 산업과 업종에 관계없이 다양한 형태로 활용되고 있으며, Fortune 500대 기업 중 400개 이상, 우수 품질관리 기업 중 70%가 도입해 사용하고 있는 것으로 알려졌다(이규환, 2008; Thornton & Rupp, 2006). 해외에서 평가센터가 신뢰로운 평가도구로 자리잡아감에 따라 많은 이론적 연구가 이어졌는데, 표 3과 같이 많은 연구들은 평가센터가 다양한 관리자의 수행(임금 상승, 승진, 훈련에서의 성공, 관리 효과성 점수)을 예측해줌을 보여주었다. 요약하자면, OAR(Overall Assessment Rating)과 성공적 관리수행의 상관관계는 .31~.43의 범위로 나타난다. 개별 평가차원(dimensions)을 이용한 준거 예측력도 OAR 만큼이나 높게 나타났으며(.25~.39), 평가차원의 가중 합산점수의 교정된 예측타당도 계수는

.45로 나타났다. 따라서 발표된 연구들에 근거할 때 평가센터가 관리자의 성공이나 수행에 대해 지니는 준거관련 타당도에는 의심의 여지가 없다(Thornton & Rupp, 2006).

그러나 우수한 준거관련 타당도에 비해, 평가센터의 구성개념 타당도는 상대적으로 취약하다고 여겨져 왔다. 특히, 평가센터의 결과가 평가차원 효과(dimension effect)가 아닌 평가과제 효과(exercise effect)를 반영한다는 견해가 우세해왔다(Bowler & Woehr, 2009). 즉, 평가센터에서 평가되는 다양한 평가차원들이 서로 변별적으로 평가되지 않고 동일한 평가과제에 속한 평가차원 끼리 유사한 점수를 받게 된다는 것이다. 반면, 이전 연구들을 리뷰하여 평가센터를 구성하는 평가차원과 평가과제의 효과를 비교한 최근 연구들에서는 다소 혼재된 결과를 보여주고 있다. Lievens과 Conway(2001)는 평가센터를 개념화하는데 평가차원과 평가과제를 조합한 혼합모형(combination model)이 가장 적합함을 확인적 요인분석을 이용해 보여줌으로써, 평가차원이 평가센터에서 중요한 부분이며 평가센터가 단지 모의활동의 나열에 지나지 않는다는 이전의 주장을 반박했으나, Lance, Lambert, Gewin, Lievens 및 Conway(2004)의 연구에서는 평가차원에 비해 평가과제 효과가 훨씬 크다는 결과를 보여주었고, Bowler와 Woehr(2006)의 연구에서는 평가과제 효과가 평가차원 효과보다 단지 근소한 차이로 크다고 주장했다. 한편, 최근 일반화가능도 이론을 이용한 Arthur, Woehr 및 Maldegen(2000)의 연구에서는 피평가자 효과, 평가차원 효과, 그리고 이들의 상호작용 효과가 전체 변량의 59%를 설명한데 반해, 평가과제 효과, 평가자 효과, 피평가자와 평가과제의 상호작용 효과는 12%에 불과한 것으로 나타났다. Jackson, Stillman 및 Atkins(2005)의 연구에서도 피평가자 효과, 평가차원 효과, 피평가자와 평가차원의 상호작용 효과가 전체변량의 36%를 설명하고, 평가과제 효과 및 평가과제와 피평가자의 상호작용 효과가 37% 차지하는 것으로 나타나, 평가과제 효과가 우세하다는 기존의 인식과 일치하지 않는 결과를 보여주었다. 한편, Bowler와 Woehr(2009)의 연구에서는 기존의 방식인 확인적 요인분석을 이용할 경우 평가과

표 3. OAR의 준거관련 타당도 검증 연구들

저자	요약
Bray & Campbell (1968)	직무수행과 .51의 상관
Byham (1970)	관리자의 수행과 .27~.64의 상관
Cohen 등 (1974)	수행과 .33, 잠재력과 .63, 승진과 .40의 상관
Borman (1982)	훈련수행과 .48의 상관
Thornton & Byham (1982)	다양한 준거(승진, 수행) 예측
Schmitt 등 (1984)	다양한 준거 예측: 직무수행(.43), 승진(.41), 성과(.31), 임금(.24)
Hunter & Hunter(1984)	직무수행과 .43, 잠재력과 .63의 상관
Gaugler 등 (1987)	(메타연구) 잠재력(.53), 수행(.36), 훈련(.35) 등 평균상관 .37
Hardison & Sackett (2004)	(메타연구) 준거관련 타당도 .31

제 효과가 우세하지만, 일반화가능도 이론을 이용할 경우 평가과제 효과는 사라지고 평가차원 효과가 강하게 나타남을 보여주었다.

일반화 가능도 이론

Cronbach 등(1963)에 의해 소개된 일반화가능도 이론은 고전검사 이론을 바탕으로 하여 발전한 신뢰도 추정 이론으로 분산분석 절차를 이용해 다양한 측정오차를 고려하여 신뢰도를 추정하는 이론이다(조재윤, 2009). 고전 검사이론이 관찰점수를 진점수(true score)와 오차점수(error score)로만 구분하는 반면, 일반화가능도 이론은 오차점수에 기여하는 다양한 원천을 구분해 내고, 각각의 원천이 갖는 상대적인 영향력을 판별해 내는 개념적인 틀과 방법론을 제공한다(김성숙, 김양분, 2001; 이규민, 2003; Brennan, 1992, 2001a; Cronbach, Glesser, Nanda, & Rajaratnam, 1972). 일반화가능도 이론에서는 이러한 원천을 '국면(facet)'이라 부르며, 각 국면의 변산성을 종합하여 관찰점수의 일반화가능도(신뢰도)를 산출한다. 한편, 일반화가능도 이론에서의 전집(universe)은 모집단(population)과 달리 관찰전집을 말하는 것으로, 모집단이 측정대상을 일컫는 반면, 일반화가능도 이론의 전집은'측정조건들'을 말하는 것이다(김성숙, 김양분, 2001). 요약하면 일반화가능도란 동일한 측정조건에서 동일한 결과가 나타날 가능성을 말하는 것이라 할 수 있다.

일반화가능도 분석은 크게 일반화 연구(Generalizability Study, G 연구)와 결정연구(Decision Study, D 연구)로 구분된다. G 연구는 각 국면의 변산성을 추정하는 연구로서 도구의 개발단계에서 많이 이용되며, D 연구는 G 연구에서 추정된 각 국면의 변산성을 이용해 전체 도구의 신뢰도를 계산하고, 적절한 신뢰도 확보를 위해 갖춰야 할 각 국면의 수준을 결정하는 연구로, 향후의 연구를 위한 제안점을 제공해준다.

G 연구(Generalizability Study)

일반화가능도 이론의 G 연구는 수행평가 결과가 동일한 조건의 다른 수행평가에 얼마나 일반화 될 수 있는가에 관심을 가지고, 측정상황에서 발생할 수 있는 다양한 오차요인(source of error)을 분산성분으로 분해하여 변량분석을 실시한 결과를 토대로 전체 평가점수에 영향을 미치는 각 국면의 변량크기를 추정하는 절차이다(남명호, 1996). 예를 들어, 수행평가의 관찰 점수는 다양한 국면 즉, 피험자에 의한 것, 평가자에 의한 것, 평가과제에 의한 것 그리고 이들의 상호작용에 의한 것으로 구분할 수 있다. G 연구가 진행되는 과정은 첫째, 측정상황의 조건에 따라 자료형태가 교차(crossed) 모형인지, 내재(nested) 모형인지 결정하여 분석모형을 설계한다. 둘째, 분산분석 결과 얻어진 각 변산원(국면)의 평균제곱(Mean Square)을 이용해 전체 관찰점수에 영향 미치는 각 변산원의 분산성분을 추정한다. 셋째, 각 국면에서 나타난 분산성분의 상대적인 크기를 비교하여 어떤 국면이 측정의 일반화 과정을 저해하는지 확인한다(김성숙, 김양분, 2001). 상대적으로 큰 분산성분을 보여주는 국면일수록 전체 점수에 큰 영향을 미치는 국면이기 때문에 신뢰도 확보를 위해서는 그러한 국면을 세분화하여 전체 점수에 미치는 영향을 감소시킬 필요가 있다.

D 연구(Decision Study)

고전검사 이론에서는 신뢰도를 향상시키기

위하여 문항 수를 증가시키는 방법을 주로 활용하지만, 일반화가능도 이론에서는 오차원(예, 평가자 수, 문항 수, 과제 수)의 영향력에 따라 각 국면 조건의 수를 다르게 조절함으로써 알맞은 신뢰도 계수를 확보하기 위한 측정 조건을 결정할 수 있다. 즉, D 연구는 G 연구의 결과로 산출된 각 국면의 분산성분 추정값을 토대로 주어진 상황에서 가장 효율적인 측정 절차를 설계할 수 있는 정보를 제공한다. D 연구에서 가장 중요한 것은 일반화전집(universe of generalizability)을 규정하는 것인데, 이는 연구목적에 맞는 특정국면을 선택하여 연구모형에 포함시키는 것을 말하며, 연구자는 연구모형에 포함된 국면의 수준을 다양하게 설계하여, 그에 따른 신뢰도 계수의 변화를 확인할 수 있다. D 연구에서는 두 가지 신뢰도 계수(일반화가능도 계수, 의존도 계수)를 제공하는데, 신뢰도 계수 산출 공식은 관찰점수분산에 대한 전집점수(관찰점수+오차점수)의 분산 비율이다. 이때 신뢰도를 산출하는데 있어 오차분산을 상대오차로 규정하는가, 절대오차로 규정하는가에 따라 일반화가능도 계수와 의존도 계수가 구분되는데, 상대오차는 '관찰대상 간의 차이파악'에 초점을 맞춘 반면, 절대오차는 '관찰대상의 행위자체의 수준'을 주시한다. 따라서 상대평가인 경우에는 상대오차를 적용한 '일반화가능도 계수'를 신뢰도 계수로 사용하고, 절대평가인 경우에는 절대오차를 적용한 '의존도 계수'를 사용한다. 상대오차를 사용하는 일반화가능도 계수는 D 연구설계에 의존한다는 것을 제외하고는 전통적인 신뢰도 계수(cronbach α)와 동일하게 정의되기 때문에(Cronbach et al, 1972) .7~.8 정도를 적절한 신뢰도 수준으로 간주한다. 한편, 의존도 계수는 상대오차 보다 큰 절대오차를 사용하기 때문에 일반화가능도 계수보다 낮게 산출되면, 따라서 .6~.8 정도를 적절한 신뢰도 수준으로 간주한다(조재윤, 2009).

연구방법

평가센터의 설계

본 연구의 평가센터는 3가지 평가과제로 구성되어 있었다. 대규모 신입사원 선발을 위해 개발된 도구인 관계로, 평가 시간을 절약하기위해 각 과제는 그 과제에서 잘 도출될 수 있는 역량을 평가하도록 설계되어 있었으며(내재설계), 구조화 면접을 담당한 평가자와 프레젠테이션 및 그룹토의를 담당한 평가자가 구분되어 있었다. 이들 평가자들은 A기업의 HR부서에서 선발하였으며, 우수 성과자로 분류된 과장급 이상의 현직자들이었다.

각 평가과제의 구성과 진행을 살펴보면, 먼저'구조화 면접(Structured Interview, SI)'에서는 각 지원자 별로 3인의 평가자가 45분간의 심층면접을 통해 3가지 평가차원 즉, 책임감, 적극성, 혁신성에 포함된 6가지 세부 활동을 평가한다. '프레젠테이션(Presentation, PT)'에서는 각 지원자가 1시간에 걸쳐 서류함검사(in basket)를 수행한 후, 그 결과를 5분간 프레젠테이션 하고 10분간 질의응답을 한다. 이 과정동안 2인의 평가자는 2가지 평가차원(정보분석력, 문제해결력)에 포함된 4가지 세부 활동을 평가한다. 마지막으로 '그룹토의(Group Discussion, GD)'에서는 6명의 지원자가 1조를 이루어 그룹토의를 30분간 진행하는데, 그룹토의의 평가자는 프레젠테이션과 동일한 평가자이며, 2가지 평가차원(대인친화력, 의사전달

력)에 포함된 4가지 세부차원을 평가한다. 모든 평가자들에 대한 참조의 틀(frame of reference)에 관한 교육은 평가에 참여하기 이전에 3박4일의 집중교육과 2일의 추가교육을 통해 이루어졌는데, 평가역량의 숙지는 물론 비디오와 실제 모의지원자에 대한 평가실습을 통해 면접스킬과 관찰 · 기록 · 분류 · 평가스킬을 향상시키기는 내용으로 진행되었다. 또한, 평가자들은 각 평가과제에서 주어진 평가차원에 대해 행동기준평정척도(BARS)에 근거한 절대평가를 하도록 교육받았다.

평가센터에 참가한 평가자의 수는 SI에 237명, PT · GD에 149명으로 평가팀별로 평균 16.3명을 평가했으며, 평가의 진행은 SI의 경우 평가자별로 하루(7시간) 9명의 지원자를 평가했으며, PT의 경우 하루 평균 18명의 지원자를 평가했고, GD의 경우 하루 평균 3회의 그룹토의를 평가했다. 전체 선발과정은 계열사 별로 1주일 이내에 이루어졌다.

자료의 구성

본 연구에 이용된 자료는 A 기업의 신입사원을 선발하기 위해 실시한 평가센터 자료이다(N=1249). 이 자료를 이용하여 평가센터의 신뢰도 및 구성개념 타당도를 분석하였다. 또한, 본 평가센터의 준거관련 타당도(동시타당도)를 검증하기 위해 재직자(관리자 105명, 일반사원 200명)를 대상으로 실시되었던 평가센터 점수와 이들의 인사고과(상사평가)자료를 이용하였다. 많은 인원과 시간이 소요되는 평가센터를 업무 중인 재직자들에게 모두 실시할 수 없는 현실적인 어려움으로 인해 재직자 평가센터는 직급별로 진행되었는데, 프레젠테이션은 평가과제의 형태는 물론 평가하는 역량(정보분석력, 문제해결력)이 관리자보다는 일선에서의 업무와 관련되었기 때문에 일반사원에게 실시되었으며, 구조화 면접은 평가역량(책임감, 적극성, 혁신성)이 업무처리보다는 인성적인 측면을 평가하기 때문에 관리자를 대상으로 실시되었다. 재직자를 대상으로 한

표 4. 평가과제 별 평가자 수 및 평균 평가횟수

평가과제	과제 별 평가자 수	전체 평가자 수	평균 평가횟수
SI	3명	237명 (79팀)	15.8명
PT/GD	2명	148명 (74팀)	16.7명
계		385명	16.3명

표 5. 재직자 대상 평가센터 자료의 구성

구분	인원	재직기간	평가과제	평가차원	준거
관리자	105	7.65년	구조화 면접	책임감 적극성 혁신성	상사평가(고객지향, 팀워크, 혁신성, 변화적응, 자기개발, 실행력, 목표설정, 의사결정, 업무처리) 직위 내 순위
일반사원	200	3.28년	프레젠테이션	정보분석력 문제해결력	

표 6. 평가역량의 구성

평가과제	평가차원	세부차원
구조화 면접(SI)	책임감	규칙준수, 결과책임
	적극성	도전정신, 성취욕구
	혁신성	창의사고, 개방사고
프레젠테이션(PT)	정보 분석력	정보수집, 정보해석
	문제 해결력	대안제시, 문제해결
그룹토의(GD)	대인 친화력	단체의식, 타인배려
	의사 전달력	적극경청, 의사전달

평가센터운영은 신입사원 선발과 동일하게 구조화 면접의 경우 3명의 평가자가 평가하였으며, 프레젠테이션의 경우 2명의 평가자가 평가하였다. 한편, 그룹토의는 운영상의 문제로 실시하지 않았다.

준거자료는 고객지향, 팀워크, 변화적응, 혁신성, 자기개발, 목표설정, 의사결정, 실행력, 업무처리 등에 대한 상사의 평가 및 직위 내 순위를 이용하였는데, 이 준거점수는 실제 A 기업이 임금, 승진 등의 기준으로 삼고 있는 인사고과 점수로 이용되고 있다. 이들 재직자의 평가센터는 2008년에 수행되었으며, 이들의 평균 재직기간은 관리자의 경우 7.65년이었으며, 일반사원의 경우 3.28년이었다.

일반화가능도 연구 설계

본 연구의 일반화 가능도 분석은 두 가지(평가과제 별, 평가과제 전체)로 진행되었다. 평가센터가 완전 교차설계 되었다면, 다시 말해 모든 평가자가, 모든 평가과제에서, 모든 평가차원을 평가하는 설계(평가자×평가과제×평가차원)라면 평가센터 점수에 미치는 평가자 효과, 평가과제 효과, 평가차원 효과를 동시에 비교할 수 있으나, 본 연구와 같이 내재설계된 평가센터의 경우에는 평가과제 별로 평가차원이 상이하기 때문에 평가과제 효과와 평가차원 효과를 동시에 확인 할 수가 없다. 따라서 평가과제 전체를 대상으로 수행한 일반화가능도 분석에서는 평가과제의 효과를 확인하고, 평가과제 별로 수행한 일반화가능도 분석에서 평가차원의 효과를 확인하고, 평가과제 별 신뢰도를 확인하였다.

평가과제 전체 대상 설계

본 연구의 평가센터는 피평가자(p)의 평가차원(d)이 평가과제(e)에 내재되어있기 때문에 p×(d:e)와 같은 분석 설계를 이용하였다.[2] 이 설계를 이용한 G 연구 결과를 이용해 평가센터를 이루고 있는 3가지 평가과제 국면(e)이 전체 평가센터점수에 미치는 영향정도를 확인할 수 있는데, 특히 피평가자와 평가과제의 상호작용(p×e) 변량이 크게 나타난다면, 이는 평가센터 점수에 평가과제 효과가 크게 영향을 미치고 있음을 보여주는 것이다. 하지만, 본 분석에서는 평가과제 효과가 크다고 해서 평가차원의 효과가 없다거나, 평가차원의 구성타당성이 존재하지 않는다고 단언할 수 없

2) 기호 : 는 왼쪽 국면이 오른쪽 국면에 내재되었다는 표시이다.

는데, 그 이유는 평가과제 마다 포함되어 있는 상이한 평가차원으로 인해 평가과제효과가 나타났을 수 있기 때문이다. 한편, 피평가자와 내재된 평가차원의 상호작용(p×d:e)은 평가차원 효과로 볼 수 없는데, 그 이유는 일반화가능도 분석에서 모든 국면이 포함된 상호작용에는 오차(error)가 혼입되어 있기 때문에 평가과제에 내재된 평가차원(d:e)의 영향을 개별적으로 확인할 수 없다. 또한, 본 연구에서는 평가과제마다 포함된 평가차원의 수 및 평가자의 수가 동일하지 않은 비균형 설계(unbalanced design)이기 때문에 비균형 설계를 분석하기 위한 urGENOVA 프로그램을 이용했으며, 이 경우 G연구만 가능하고, D연구는 실시할 수 없는 관계로 이 설계모형에서는 신뢰도가 산출되지 않는다.

평가과제 별 설계

본 평가센터 자료는 평가과제에 따라 평가자와 평가차원이 다르지만, 각 평가과제 별로 배정된 평가차원(d)에서는 동일한 평가자(r)들이 평가하기 때문에, 평가과제 내에서는 교차된 설계 즉, p×r×d 설계를 이용하였다. 이 교차설계(p×r×d)에 따른 G 연구결과를 이용해 각 평가과제의 점수에 영향을 미치는 피평가자(p), 평가자(r), 평가차원(d)의 주효과는 물론, 피평가자와 평가자의 상호작용(p×r), 피평가자와 평가차원의 상호작용(p×d), 평가자와 평가차원의 상호작용(r×d) 효과의 크기를 비교할 수 있다. 여기서 평가자(r)의 주효과나 피평가자와 평가자의 상호작용 효과(p×r)가 크게 나타날 경우, 이는 평가자에 따라 평가과제 점수가 달라진다는 증거이므로 평가자 편향(rater bias)이 존재한다는 것이며, 따라서 추가적인 평가자 참조의 틀 교육이나, 객관적 평가기준이 요구된다고 할 수 있을 것이다. 한편, 피평가자(p)의 주효과나 평가차원(d)의 주효과 특히, 이들의 상호작용 효과(p×d)가 크게 나타난다면, 이는 동일한 과제 내에서 피평가자의 평가차원이 차별적으로 평가되고 있음을 보여주는 것이므로 구성개념 타당도의 근거로 볼 수 있다. 또한 이 설계(p×r×d)를 이용한 D 연구의 결과는 이러한 설계의 평가과제의 신뢰도를 보여주며, 평가과제 결과가 일반화되기 위해 필요한 각 국면의 수준 수를 결정할 수 있게 해준다.

확인적 요인분석 모형

평가센터의 점수는 피평가자가 평가과제를 수행하면서 보여주는 평가차원 관련 행동이 반영된 것이다. 이 평가점수가 평가차원과는 무관하게 평가과제에 따라 다르게 나타난다면 이는 평가과제 효과가 우세한 것이라 할 수 있다. 반대로 평가점수가 평가과제와는 무관하게 평가차원 별로 다르게 나타난다면 평가차원의 변별성을 보여주는 것이라 할 수 있다. 확인적 요인분석은 평가점수의 요인구조가 평가과제 요인, 평가차원 요인 중 어떤 요인으로 설명되는지를 살펴보기 위한 것이다. 본 연구에서는 동일한 평가차원이 여러 평가과제에서 평가되지 않기 때문에, 평가과제별로 확인적 요인분석을 실시하였다. 검증모형은 3가지로, 먼저 모형 1은 평가과제 내에서 측정되는 개별적 역량들의 효과를 무시하고 과제효과만을 잠재변인으로 설정한 일요인 모형으로서 평가과제 효과를 확인하는 모형이다. 모형 2는 각 평가과제에서 평가되는 평가차원을 잠재변인으로 설정한 다요인 모형으로서 평가차원의 변별성을 확인하는 모형이다. Gibbons,

Rupp, Baldwin 및 Holub(2005)의 연구에서도 평가차원에 포함된 세부평가차원이 개별적으로 평가됨을 입증함으로써 평가센터의 구성개념 타당도 증거를 제시한바 있다. 모형 3은 다특성다측정(MTMM) 검증 모델을 응용한 것으로 평가과제와 평가차원을 모두 가정한 모형이다. 한편, 이와 같은 분석이 평가과제 내의 평가차원의 변별성은 확인할 수 있지만, 평가센터 전반에서 나타나는 평가과제의 효과를 확인하기에는 부적합할 수 있다는 지적이 있을 수 있다. 따라서 평가센터의 3가지 평가과제(3 Exercise)와 7가지 평가차원(7 Dimension) 전체를 대상으로 확인적 요인분석을 설계하여, 3가지 평가과제만으로 구성된 모형 4, 7가지 평가차원 만으로 구성된 모형 5, 그리고 3가지 평가과제와 7가지 평가차원을 모두 포함한 모형 6의 적합도를 비교하였다. 확인적 요인분석 결과, 평가과제 모형(모형 1, 모형 4)이 우수한 것으로 판단될 경우, 평가차원 도출을 위한 직무분석이나 역량모델링과 같은 과정의 중요성이 약해진다고 할 수 있다. 반면, 평가차원 모형(모형 2, 모형 5)이나 혼합모형(모형 3, 모형 6)이 우수한 것으로 판단될 경우에는 피평가자들이 평가차원에 따라 차별적으로 평가되고 있다는 것이므로, 이는 평가센터를 구성하는 평가차원의 구성개념 타당도를 지지해주는 증거가 된다.

모형에 포함되어 있는 측정변인들은 다수의 평가자들이 평가한 평가차원 점수들이며, 동일한 평가차원들의 오차 간 상관을 설정하였다.

분석도구

일반화가능도 분석을 위해서는 Crick과 Brennan(1983)이 개발한 컴퓨터프로그램인

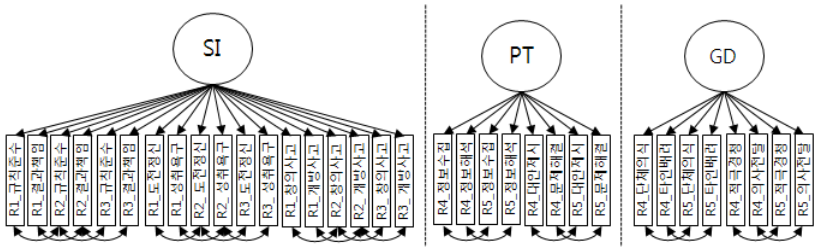

그림 1. 모형 1: 과제별 평가과제모형

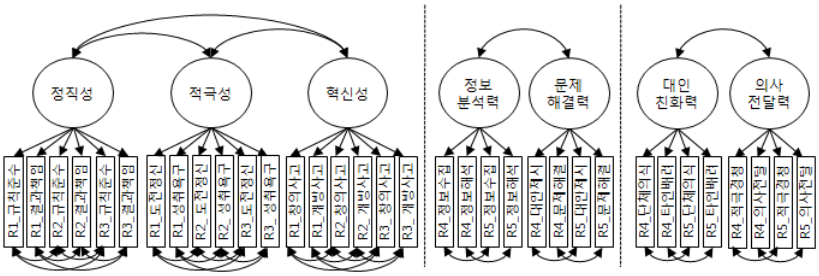

그림 2. 모형 2: 과제별 평가차원 모형

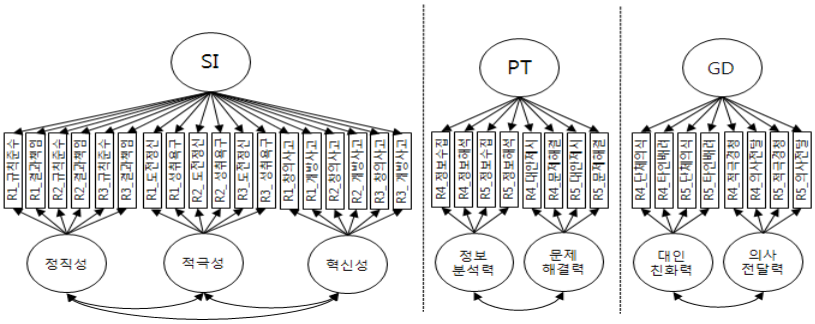

그림 3. 모형 3: 과제별 혼합 모형

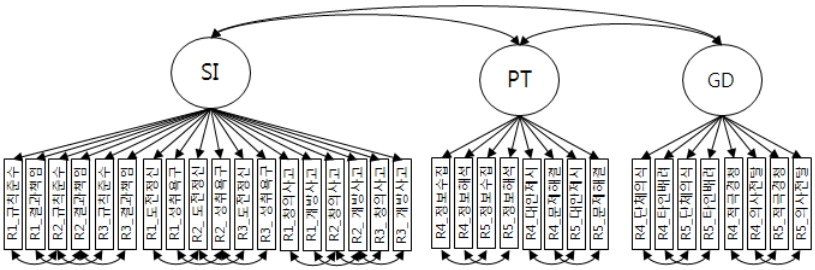

그림 4. 모형 4: 전체 평가점수의 평가과제 모형

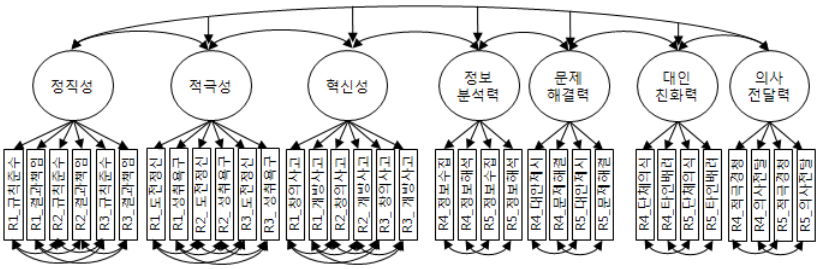

그림 5. 모형 5: 전체 평가점수의 평가차원 모형

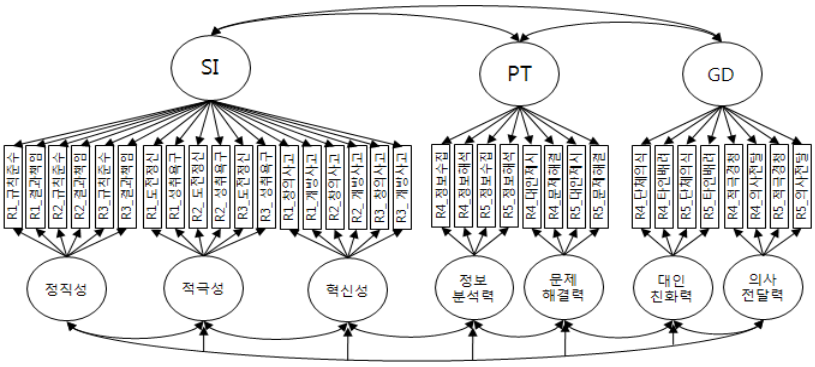

그림 6. 모형 6: 전체 평가점수의 혼합 모형

GENOVA 3.1(Crick & Brennan, 1983)을 이용하였으며, 확인적 요인분석은 AMOS 16.0을 이용하였고, 기타 통계분석은 SPSS 16.0을 이용하였다.

연구결과

신뢰도 검증

평가자 국면(R)과 평가차원 국면(D)을 고려한 p×R×D[3] 설계를 이용해 본 평가센터의 신뢰도를 산출하였다. 일반화가능도 분석의 '결정연구(D 연구)'는 두 가지 신뢰도 계수를 산출해준다. 첫째, 일반화가능도(generalizability) 계수는 상대평가의 신뢰도 계수이며, 0.7 이상을 비교적 신뢰로운 수준으로 판단한다. 둘째, 의존도(dependability) 계수는 절대평가의 신뢰도 계수이며, 0.6 이상인 경우에 비교적 신뢰롭다고 판단한다. 본 연구의 평가센터는 평가자 훈련과정을 통해 평가차원 별로 BARS (Behavioral Anchored Rating Scale)에 근거한 절대평가를 하도록 했기 때문에, 절대평가 신뢰도 계수인 의존도 계수를 이용해 신뢰도를 판단하였다. 한편, 평가센터와 같은 수행평가는 다수의 평가자와 다수의 평가차원을 특징으로 하기 때문에, 이러한 평가방식을 실제 적용하는 과정에서 시간적, 경제적 등 다양한 실행상의 이유로 평가자 수, 평가차원의 수 등의 제한으로부터 자유로울 수 없다. 따라서 D 연구의 분석프로그램에 평가자 수를 2명부터 4명까지, 평가차원의 수를 2개에서 6개까지 조절하여 투입함으로써 신뢰도 추정치의 변화를 분석하였다.

표 7은 각 평가과제의 신뢰도를 보여주며, 그림 7은 평가차원의 수와 평가자수를 조정했을 때 신뢰도의 변화추이를 보여주고 있다.

구조화 면접(SI)의 신뢰도

본 연구의 구조화 면접 조건(평가자 3인, 평가차원 3개)에서, 일반화가능도 계수와 의존도 계수는 .62555와 .61681로 의존도 계수의 경우 기준을 만족시키고 있었다. 신뢰도를 향상시키기 위해서는 평가자 수를 고정시킨 상태에서 평가차원 수를 5개로 증가시킬 경우 두 신뢰도 기준 모두 0.7이상으로 향상되는 것으로 나타났으며, 평가차원 수를 3개로 고정할 경우 평가자의 수를 4명으로 증가시켜도 신뢰도가 0.7 기준을 만족시키지 못하는 것으로 나타났다. 한편, 평가자의 수를 2명으로 줄이더라도 평가차원의 수를 4개 이상으로 증가시키면 현재 수준 이상의 신뢰도를 얻을 수 있는 것으로 나타났다. 결과적으로 구조화 면접의 신뢰도를 보다 향상시키기 위해서는 평가자의 수 보다 평가차원의 수를 증가시키는 것이 효과적이라고 할 수 있으나, 평가도구의 효율성을 고려할 때 현재의 수준을 유지하는 것이 바람직하다고 할 수 있다.

프레젠테이션(PT)의 신뢰도

평가센터는 복수의 평가자가 복수의 평가차원에 대해 평가를 진행하기 때문에, 본 연구의 프레젠테이션 조건(평가자 2인, 평가차원 2개)은 평가센터의 최소 구성이라고 할 수 있다. 프레젠테이션의 결정연구 결과를 보면, 일반화가능도 계수와 의존도 계수가 .81448과 .81385로 모두 높은 신뢰도를 보여주었다. 평가자의 수를 현재 보다 많은 3명 이상으로 조

3) 일반화가능도 이론에서는 피평가자를 제외한 다른 국면에 대해, G연구에서는 소문자로, D연구에서는 대문자로 표시하도록 하고 있다.

표 7. 평가과제 별 국면구성 및 신뢰도

	SI	PT	GD
평가자 수	3	2	2
평가차원 수	3	2	2
일반화가능도 계수	.62555	.81448	.76919
의존도 계수	.61681	.81385	.76847

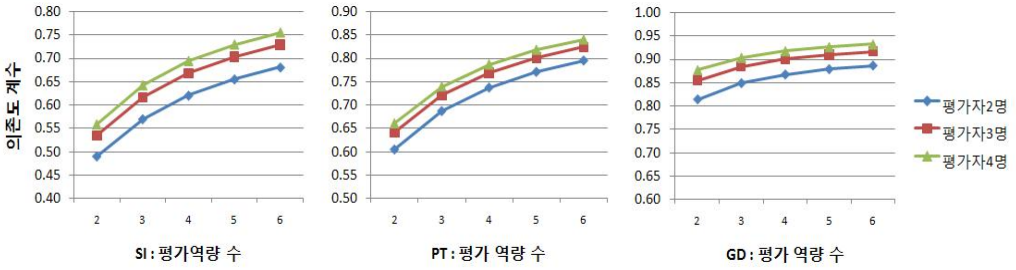

그림 7. 국면의 조정에 따른 신뢰도(의존도계수)의 변화 추이

정하거나, 평가차원의 수를 현재 보다 많은 3개 이상으로 조정할 경우 오차가 감소하며 신뢰도가 향상되지만, 현재와 같은 최소한의 구성으로도 만족할 수준이기 때문에 평가의 효율성을 고려한다면 평자자와 평가차원의 수를 증가시킬 필요성이 없는 것으로 나타났다.

그룹토의(GD)의 신뢰도

본 연구의 그룹토의는 프레젠테이션과 마찬가지로 최소 구성으로 설계되었다. 그룹토의는 피평가자 6명이 한 조로 진행되었으며, 2명의 평가자가 2가지 평가차원을 평가하도록 설계되어 있었다. 그룹토의의 결정연구 결과를 보면, 일반화가능도 계수와 의존도 계수가 .76919와 .76847로 두 신뢰도 모두 신뢰로운 수준이었다. 평가자의 수나 평가차원의 수를 늘릴 경우 향상되는 신뢰도의 수준을 보면, 평가자의 경우 3명 이상으로 증가시킬 경우 신뢰도의 증가폭이 작아졌으며, 평차차원의 경우에도 평가차원의 수를 4개 이상으로 증가시킬 경우에는 신뢰도 증가폭이 작아졌다. 따라서 현재의 평가자와 평가차원 수준을 유지해도 신뢰도에는 문제가 없으며, 증가시킬 경우 평가자는 3명을, 평가차원은 4개를 넘지 않는 선에서 증가시키는 것이 효율적이라고 할 수 있다.

타당도 검증

G 연구(구성개념 타당도)

평가차원(d)이 내재되어 있는 세 가지 평가과제(e)가 피평가자(p)의 평가결과에 미치는 영향을 확인하기 위해, 평가차원이 내재된 평가과제 국면(d:e)을 일반화연구 설계에 포함시켜 (p×(d:e)설계) 평가결과에 미치는 영향력 즉, 평가과제 효과(exercise effect)를 확인하였다. 먼저, 평가과제 국면을 살펴보면, 평가과제의 주효

과(e)는 12.8%로 어느 정도 영향력을 보여주었으며, 특히 피평가자와 평가과제의 상호작용 효과(p×e)는 37.0%로 높게 나타났다. 이는 전체 평가센터에서 피평가자가 평가과제에 따라 상이한 평가를 받는 평가과제 효과를 보여주는 것이다. 한편, 피평가자와 내재된 평가차원의 상호작용 효과(p×d:e)가 24.8%로 나타나긴 했으나 일반화가능도 분석에서는 모든 국면의 상호작용(p×d:e))에는 오차(error)가 혼입되는 관계로 이를 평가차원의 효과로 해석할 수 없다. 평가과제 별로 피평가자(p)의 평가점수에 영향을 미치는 평가자(r) 국면과 평가차원(d) 국면의 영향을 확인하기 위해 p×r×d 설계에 따른 일반화연구를 실시한 결과, 그 결과 구조화 면접에서 평가자 국면의 주효과(r)는 0.1%로 거의 나타나지 않았으며, 피평가자와 평가자의 상호작용(p×r)은 9.9%로 나타났다. 평가차원 국면의 주효과(d)는 1.8%로 매우 약하게 나타났으나, 피평가자와 평가차원의 상호작용(p×d)은 27.2%로 높게 나타났다. 한편, 평가자 국면과 평가차원 국면의 상호작용(r×d)은 0.1%로 거의 나타나지 않았다. 프레젠테이션에서는 피평가자(p) 국면이 60.4%로 매우 높게 나타났다. 이는 피평가자의 전반적인 특성이 프레젠테이션 점수의 매우 큰 부분을 설명한다는 것을 의미한다. 이러한 경우 평가자 국면과 평가차원 국면의 영향력은 상대적으로 약하게 나타날 것이다. 결과적으로, 평가자 국면의 주효과(r)는 전혀 나타나지 않았으며, 피평가자와 평가자의 상호작용(p×r)은 9.3%로 약하게 존재하는 것으로 나타났다. 평가차원 국면에서는 주효과(d)가 0.1%로 거의 없는 것으로 나타났으나, 피평가자와 평가차원의 상호작용(p×d)은 6.2%로 약하게 존재하는 것으로 나타났다. 한편, 평가자 국면과 평가차원 국면의 상호작용(r×d)은 전혀 나타나지 않았다. 그룹토의에서도 마찬가지로 피평가자(p) 국면이 52.9%로 매우 높게 나타났다. 이는 프레젠테이션과 마찬가지로 그룹토의 역시 피평가자의 전반적인 특성이 그룹토의 점수의 매우 큰 부분을 설명한다는 것을 의미한다. 평가자 국면을 살펴보면, 평가자의 주효과(r)는 0.1%로 거의 나타나지 않았으나, 피평가자와 평가자의 상호작용(p×r)은 8.9%로 약하게 존재하는 것으로 나타났다. 평가차원 국면에서는 주효과(d)가 전혀 나타나지 않았으나, 피평가자와 평가차원의 상호작용(p×d)은 7.7% 약하게 나타났다. 평가자 국면과 평가차원 국면의 상호작용(r×d)은 전혀 나타나지 않았다.

결과적으로 세 평가과제 모두에서 평가자(r)와 관련된 효과를 살펴보면, 주효과(r)는 거의 나타나지 않아 평가자에 따른 편향은 없었음을 보여주었고(0.0~0.1%), 평가차원과 평가자(r×d)도 거의 없는 것으로 나타나 평가차원 따라 평가자의 평가가 달라지지 않았음을 보여주었다(0.0%~0.1%). 그러나 피평가자와 평가자의 상호작용(p×r)은 8.9~9.9%로 동일한 피평가자가 평가자들로부터 어느 정도 다른 평가를 받았음을 보여주었다. 한편, 평가차원의 효과를 보여주는 피평가자와 평가차원의 상호작용(p×d)은 6.1~27.2%로 나타났는데 특히, 27.2%로 나타난 구조화 면접에서 피평가자들이 평가차원에 따라 차별적인 평가를 받았음을 보여주었다. 그 밖에 프레젠테이션과 그룹토의에서는 각각 6.2%와 7.7%로 나타나 구조화 면접에서 비해 평가차원의 효과가 약한 것으로 나타났다. 또한, 프레젠테이션과 그룹토의는 피평가자 변량(p)이 각각 60.4%와 52.9%로 매우 높게 나타나, 구조화 면접(27.0%)에 비해 피평가자의 전반적 개인특성이 평가에

표 8. 전체 검사의 오차원 분석결과 요약(G연구)

분산성분	MS	분산 추정치	%
피평가자(p)	0.93679	0.08028	24.6
평가과제(e)	123.4151	0.04193	12.8
평가차원:평가과제(d:e)	3.43141	0.00268	0.8
pe	0.35755	0.12093	37.0
pd:e	0.08115	0.08115	24.8
계		0.32697	100.0

표 9. 평가과제 별 오차원(source of error) 분석결과 요약(G study)

분산성분	SI		PT		GD	
	분산추정치	%	분산추정치	%	분산추정치	%
피평가자(p)	0.09930	27.0	0.31724	60.4	0.21342	52.9
평가자(r)	0.00019	0.1	0.00013	0.0	0.00032	0.1
평가차원(d)	0.00646	1.8	0.00045	0.1	0.00016	0.0
pr	0.03655	9.9	0.04911	9.3	0.03570	8.9
pd	0.10003	27.2	0.03238	6.2	0.03107	7.7
rd	0.00025	0.1	0.00002	0.0	0.00010	0.0
prd	0.12527	34.0	0.12608	24.0	0.12260	30.4
계	0.36805	100.0	0.52540	100.0	0.40337	100.0

매우 강하게 반영되고 있음을 보여주었다. 마지막으로 잔여변량을 말하는 모든 국면의 상호작용(p×r×d)은 24.0~34.0%로 높게 나타났는데, 잔여변량 성분이 전체 점수 분산에서 차지하는 비중이 큰 것은 일반화가능도 이론 연구에서 일반적인 현상으로, 잔여변량 성분에는 모형에서 구성된 요인들로 설명되지 않는 분산성분(오차) 부분이 모두 포함되기 때문이다(강애남, 이규민, 2006).

확인적 요인분석(구성개념 타당도)

평가센터 결과점수를 측정변인으로 삼고, 평가차원과 평가과제를 잠재변인으로 하는 구조방정식 모형을 검증하여, 평가센터 결과에서 평가차원의 구성개념 타당도를 확인하였다. 먼저, 평가과제만으로 구성된 모형 1과 평가차원만으로 구성된 모형 2, 그리고 평가과제와 평가차원을 모두 포함한 모형 3을 평가과제 별로 비교하였다. 또한, 평가과제 간의 상호작용을 고려할 경우 평가차원 효과보다 평가과제의 효과가 부각될 수도 있음을 감안하여, 본 평가센터에서 도출된 모든 평가점수를 서로 상관된 3가지 평가과제의 효과로 설정한 모형 4와 7가지 평가차원으로 설정한 모형 5,

그리고 두 효과 모두로 설정한 모형 6의 적합도를 비교하였다.

평가과제 별 확인적 요인분석 결과, 세 가지 평가과제(SI, GD, PT) 모두에서 평가과제(모형 1)나 평가차원(모형 2)만으로 구성된 모형은 비슷한 수준의 낮은 적합도를 보여주었으나, 혼합모형(모형 3)의 적합도는 수용가능한 수준의 우수한 적합도를 보여주었다. 즉, 평가과제에서 도출된 평가자들의 평가점수들이 과제의 특성이나 역량의 특성만을 반영하는 것이 아니라, 평가과제와 평가역량 모두의 특성을 반영한다는 것을 의미하며, 이는 평가점수에 평가과제 효과(exercise effect)와 평가차원 효과(dimension effect)가 모두 존재한다는 것을 말한다. 평가센터 전체에 대한 확인적 요인분석 결과에서도 마찬가지로, 평가과제와 평가차원을 모두 가정한 모형 6의 모형 적합도가 가장 우수한 것으로 나타났다. 즉, 평가센터 전체 점수를 평가과제의 특성이나 평가역량의 특성으로 만으로 해석하는 것보다 둘 모두를 반영하는 것으로 보아야 하며, 이는 전체 평가센터 평가점수에서도 평가과제 효과와 평가차원 효과가 모두 존재한다는 것을 말한다. 이러한 결과는 평가센터 점수가 개별 평가과제에서 피평가자가 보여주는 일반적인 인상효과나 후광효과만을 반영하는 것이 아니라, 평가차원에 대한 차별적인 평가가 이루어지고 있음을 보여주는 것으로, 평가센터 점수의 구성개념 타당도를 지지하는 증거로 볼 수 있다.

표 10. 평가과제 별 구성타당성 검증을 위한 확인적 요인분석

모형		Chi	df	CFI	TLI	RMSEA
SI	모형1(1E)	2111.16	117	.856	.812	.111
	모형2(3D)	1830.65	114	.863	.817	.110
	모형3(1E3D)	626.84	96	.959	.935	.065
PT	모형1(1E)	487.03	16	.940	.896	.154
	모형2(2D)	372.26	15	.955	.916	.138
	모형3(1E2D)	22.34	7	.998	.992	.042
GD	모형1(1E)	409.03	16	.934	.885	.140
	모형2(2D)	316.87	15	.949	.906	.127
	모형3(1E2D)	24.56	7	.997	.988	.045

표 11. 전체 평가센터의 구성타당성 검증을 위한 확인적 요인분석

모형	Chi	df	CFI	TLI	RMSEA
모형 4(3E)	3570.02	498	.890	.876	.070
모형 5(7D)	3166.80	480	.904	.887	.067
모형 6(3E7D)	1164.34	443	.974	.967	.036

상관관계 분석(준거관련 타당도)

본 평가도구의 준거관련 타당도(동시타당도)를 검증하기 위해 재직자(관리자, 일반사원)를 대상으로 본 평가센터를 실시한 후 이들의 평가센터 점수와 상사평가 점수의 상관관계를 분석하였다. 준거점수인 상사평가 점수는 고객지향, 팀워크, 변화적응, 혁신성, 자기개발, 목표설정, 의사결정, 실행력, 업무처리 등에 대한 상사평가 및 직위 내 순위였다.

먼저, 관리자를 대상으로 한 구조화 면접의 평가점수와 준거의 상관관계를 살펴보면, 도전정신과 성취욕구를 평가하는 '적극성'의 경우, 모든 상사평가 준거와 상관을 보여주었으며, 특히 업무처리(r=.350, p<.01)와 실행력(r=.327, p<.01)에서 높은 상관을 보여주었다. 이는 '적극성'이 다른 평가차원에 비해 실제 직무수행에 대해 폭넓은 준거예측력이 있음을 보여주는 것이라 하겠다. 한편, 규칙준수 및 결과책임을 평가하는 '책임감'의 경우 변화적응을 제외한 나머지 역량에서는 유의한 상관이 나타나 '적극성'만큼이나 넓은 준거예측력을 지닌다고 할 수 있는데, 특히 목표설정(r=.366, p<.01)과 팀워크(r=.362, p<.01), 업무처리(r=.355, p<.01)에서 높은 상관을 보여주었다. 그러나 변화적응과는 상관을 보여주지 않았는데, 이는 '책임감'의 평가내용인 약속 및 규칙준수, 결과책임이 조직 내의 업무수행과 관련성이 높은데 반해, 개인의 환경변화에 대한 적응과는 관련성이 낮기 때문으로 판단된다. 창의사고와 개방사고를 평가하는 '혁신성'의 경우 실행력(r=.291, p<.01), 목표설정(r=.240, p<.05), 변화적응(r=.225, p<.05), 혁신성(r=.224, p<.05) 등과 유의한 상관관계를 보여준 반면, 고객지향, 팀워크, 업무처리와는 유의한 상관관계가 나타나지 않아, 구조화 면접에서 평가하는 차원들 중 가장 준거예측력이 낮게 나타났다. 이는 혁신성의 평가내용인 창의와 개방적 사고가 융통성 있는 대처를 필

표 12. 재직자 평가센터점수와 준거(상사평가)의 상관관계

준거		SI (관리자)			PT (일반사원)	
		책임감	적극성	혁신성	정보 분석력	문제 해결력
상사평가	고객지향	.220*	.271**	.097	-.097	-.073
	팀워크	.362**	.299**	-.026	-.070	-.129
	업무처리	.355**	.350**	.166	.139*	.143*
	의사결정	.289**	.209*	.199*	-.051	-.018
	실행력	.306**	.327**	.291**	.031	.100
	목표설정	.366**	.261**	.240*	.117	.036
	자기개발	.229*	.244*	.224*	.015	.068
	혁신성	.304**	.268**	.224*	.100	.076
	변화적응	.191	.268**	.225*	-.009	-.002
직위 내 순위		.517**	.548**	.346**	.140*	.150*

* p<.05 ** p<.01

요로 하는 변화적응, 혁신, 실행 등의 준거와 관련된데 반해 대인 지향적이거나 업무지향적인 고객지향, 팀워크 및 업무처리 준거와는 관련성이 없다는 것으로 해석할 수 있다.

일반사원을 대상으로 한 프레젠테이션의 평가점수와 준거의 상관관계에서는, 정보수집 및 정보해석을 평가하는 '정보분석력'의 경우 업무처리 준거와 유일한 상관관계를 보여주었고(r=.139, p<.05), 대안제시와 문제해결을 평가하는 '문제해결력' 역시 업무처리 준거에서만 상관관계를 보여주었다(r=.143, p<.05). 따라서 프레젠테이션에서 평가하는 '정보분석력'이나'문제해결력'은 실제 업무처리에서만 관련성을 보여주는 과제지향적인 평가차원으로 볼 수 있었다.

한편, 직위 내 순위와의 상관에서는 구조화 면접과 프레젠테이션에서 평가하는 모든 역량이 통계적으로 유의한 상관관계를 보여주었는데, 특히 구조화 면접에서 평가하는 책임감(r=.517, p<.01), 적극성(r=.548, p<.01), 혁신성(r=.346, p<.01)이 프레젠테이션에서 평가하는 정보분석력(r=.140, p<.05)과 문제해결력(r=.150, p<.01)보다 높은 상관관계를 보여주었다.

논 의

평가센터는 다수의 평가자가 다수의 평가방식을 통해 다수의 평가차원에 대해 평가를 수행한다. 평가센터는 다른 여러 평가방식에 비해 준거관련 타당도 측면에서의 우수성이 증명되어 왔다. 하지만, 평가센터의 과제들은 전통적인 자기보고식 지필검사와 달리 다수의 평가자가 피평가자의 수행을 평가하기 때문에, 피평가자의 능력 이외에도 평가자, 평가과제, 평가과제의 수 등의 다양한 요소들과 이러한 요소들의 상호작용의 영향을 받을 수 있다. 때문에 준거관련 타당도에 비해 상대적으로 구성타당도에 대해서는 비관적인 견해가 지배적이었으며, 구성타당도에 대한 연구결과들 또한 일관적인 결과를 보여주지 않았다. 따라서 평가센터와 같은 수행평가는 그 활용에 앞서 결과의 신뢰도와 타당도 검증이 선행되어야 할 필요가 있다.

본 연구는 기업의 신입사원 채용에 적용된 내재설계 평가센터의 신뢰도와 타당도를 확인했다. 신뢰도를 확인하기 위한 분석으로 일반화가능도 이론의 결정연구(D study)를 이용해 신뢰도 계수를 추정한 결과, 각 평가과제 별 평균 의존도 계수가 구조화 면접이 .67, 프레젠테이션이 .81, 그룹토의가 .72로 나타나 절대평가에서 기준으로 보는 의존도 계수의 신뢰도 기준인 .60 보다 높은 값을 보여주었다. 따라서 각 평가과제가 현재 포함하고 있는 측정조건의 신뢰도는 양호한 것을 알 수 있었으며, 각 평가과제의 신뢰도를 향상시키기 위해 추가적인 평가자나 평가차원의 증설은 요구되지 않는다고 할 수 있다.

구성개념 타당도를 확인하기 위한 첫 번째 분석으로 일반화가능도 이론의 일반화 연구(G study)를 통해 평가과제 효과를 살펴본 결과, 피평가자와 평가과제 국면의 상호작용 가장 높게 나타나(37.0%), 피평가자에 따라 우수한 수행을 보이는 과제가 존재함을 보여주었다. 이러한 결과는 평가센터에 평가과제 효과가 존재하기 때문으로 생각할 수 있으나, 평가차원이 평가과제에 내재된 설계의 경우 특정 평가과제 점수가 그 평가과제에만 포함된 평가차원들의 점수로 형성되는 것이므로, 평가차원의 영향력을 완전히 배제할 수는 없을 것이

다. 한편, 각 평가과제 별로 평가차원 효과를 살펴본 결과, 각 평가과제 내에서 평가차원의 효과가 어느 정도 타나나고 있었으며(6.2~27.2%), 특히 구조화 면접에서 평가차원 효과가 강하게 나타났는데, 이러한 결과는 구조화 면접의 경우 3명의 평가자가 지원자에게 각각의 역량과 관련된 질문을 하고 그에 대한 지원자의 답변을 역량별로 평가하는 과정으로 진행되었기 때문에 역량들이 차별적인 평가되기가 수월한데 반해, 프레젠테이션은 일정시간동안 지원자의 발표를 듣고 추가적인 질의응답을 통해 평가하고 그룹토의는 지원자들이 수행하는 토론의 관찰한 후에 질의응답 없이 평가하기 때문에 역량들이 차별적으로 평가되기 어렵기 때문인 것으로 생각된다.

구성개념 타당도를 확인하기 위한 두 번째 분석으로 확인적 요인분석이 실시되었다. 그 결과, 평가과제나 평가차원만을 고려한 모형보다, 양자 모두를 고려한 혼합모형의 적합도가 가장 우수하게 나타났다. 이는 평가과제효과가 존재하는 것은 사실이나 평가차원의 효과 또한 존재하기 때문에 평가차원을 무시하는 것은 옳지 않다는 최근의 평가센터 구성개념 타당도 연구결과들(Lievens et al, 2001; Bowler et al, 2006, 2008)과 맥을 함께 하는 결과이다.

마지막으로 평가센터 평가점수와 재직자의 준거점수(상사평가)의 상관관계를 분석한 결과를 종합하면, 구조화 면접에서 평가한 적극성과 책임감이 보편적으로 준거관련성이 높게 나타났으며, 다음으로는 혁신성이 대인지향적 준거(고객지향, 팀워크)나 업무처리를 제외한 나머지 준거들과 관련성을 보여주었다. 프레젠테이션에서 평가한 정보분석력, 문제해결력의 경우 업무처리 준거에서만 유의한 상관을 보여주었다. 한편, 모든 평가차원이 직위 내 순위와 유의한 상관관계를 보여주었다. 특정 평가차원이 특정 준거와 높은 상관을 보이며, 이러한 평가차원과 준거 간에는 내용의 유사성이 존재했다는 점은 평가차원의 구성개념 타당도의 증거로도 볼 수 있을 것이다. 한편, 평가과제 별로 해석하면, 구조화 면접이 프레젠테이션 보다 평가차원의 준거예측력이 우수하다고 볼 수 있는데, 이는 G연구 결과에서 구조화 면접에서의 평가차원 관련변량(d, p×d)이 29.1%인데 반해, 프레젠테이션의 평가차원 관련변량(d, p×d)은 6.3%로, 구조화 면접에서 평가차원이 보다 변별적으로 평가되었기 때문인 것으로 판단된다. 하지만, 본 연구에서 제시한 준거관련 타당도의 증거는 평가과제 별 평가대상 재직자의 이질성 및 프레젠테이션의 낮은 준거관련성으로 인해 제한적인 범위에서 해석되어야 할 것이다.

결과적으로 내재설계 평가센터는 양호한 신뢰도를 보여주었으며, 구성개념 타당도 측면에서는 상당한 수준의 평가과제 효과가 발견되었으나, 평가차원 간의 변별성 증거 역시 확인되었다. 따라서 평가차원을 무시한 채 평가과제나 역할위주로 평가센터를 구성해야 한다는 주장은 적절치 않음을 확인할 수 있었으며, 내재설계의 평가센터에서도 직무분석이나 역량모델 구축 등을 통해 평가차원을 선별하고, 평가차원 별로 평가를 진행하는 절차는 평가의 정확성을 높이는 의미 있는 과정이라고 할 것이다. 또한, 내재설계(nested design)된 평가센터의 결과를 해석할 때는 평가차원 점수를 그대로 해석하기보다 평가과제의 효과를 감안해야할 것이다. 따라서 각 평가차원의 점수를 바탕으로 근로자를 평가하거나 근로자 인적자원개발 계획을 수립하려는 진단목적의

평가센터나 개발목적의 평가센터에서는 내재설계가 아닌 교차설계(crossed design) 즉, 동일한 평가차원이 다수의 평가과제에서 반복 측정되는 평가센터를 설계하여 활용해야 할 것이다. 또한 인사선발 목적의 평가센터에서도 관리가능한 적절한 인원을 대상으로 할 경우에도 교차설계가 보다 타당할 것이며, 본 연구와 같이 대규모 인원 선발에 사용된 내재설계 평가센터의 결과는 인사선발 목적에 국한하여 활용하는 것이 적절하다고 할 수 있다.

본 연구가 가지는 시사점은 첫째, 점차 확대되고 있는 평가센터의 활용에 따라 등장한 다양한 설계의 평가센터 중 일반 신입사원을 선발하기 위해 사용되고 있는 내재설계 평가센터의 신뢰도 및 타당도를 검증함으로써 향후 이와 같은 평가센터가 활용될 수 있는 심리측정학적 근거를 마련하였다는 점을 들 수 있다. 둘째, 평가센터의 신뢰도를 확보하기 위한 평가국면들의 수준을 제시함으로써, 평가센터의 측정절차를 설계하는 과정에서 마주치는 시간적, 경제적, 실용적 제한점에 대한 판단의 기준을 제공하였다는 점이다. 비록 본 연구의 제안하는 국면의 수준들이 본 연구에서와 같은 평가센터 구조에 대한 것이지만, 다양한 설계의 평가센터에 대한 연구들이 축적되어간다면 평가센터 설계에 대한 유용한 지침이 형성될 것으로 생각한다.

한편, 본 연구는 몇 가지 사항에서 제한점을 가진다. 첫째, 재직자를 대상으로 한 평가차원과 준거의 상관관계 분석에서 평가의 적절성을 고려하여 관리자는 구조화 면접을 일반사원은 프레젠테이션을 수행하였으나, 그로 인해 평가차원과 준거의 상관관계가 평가과제의 효과로 해석될 수도 있고, 피평가자 직급의 효과로 해석될 수도 있는 여지가 있었다. 따라서 상이한 평가과제에 포함된 상이한 평가차원 간의 준거예측력 비교를 위해서는 모든 평가과제에서 모든 직급의 평가자료를 수집해야 할 것이며, 적어도 평가과제 간 피평가자의 직급을 통일할 필요가 있었다. 둘째, 내재설계 평가센터의 자료만을 분석 비교함으로써 내재설계 평가센터와 교차설계 평가센터의 유사점과 차이점을 직접 비교할 수 없었다는 제한점이 있다. 셋째, 본 평가센터의 자료는 한 기업에 해당하는 자료이기 때문에 본 연구의 결과를 다른 모든 내재설계 평가센터로 일반화시키는 것은 무리가 있으며, 향후 내재설계 평가센터를 이용하는 여러 기업의 평가센터 연구결과가 축적되어야 할 것이다.

참고문헌

강애남, 이규민 (2006). 학생들의 동료평가를 활용한 수행평가 결과의 일반화가능도 분석. 교육평가연구, 19, 107-121.

김성숙, 김양분 (2001). 일반화가능도 이론. 서울: 교육과학사.

남명호 (1996). 수행평가에 있어서 일반화가능도 이론의 활용. 교육평가연구, 9, 73-93.

남명호 (2002). 수행평가: 기술적 측면. 서울: 교육과학사.

성태제 (2002). 현대교육평가. 서울: 학지사.

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

이규환 (2008). 인재선발과 AC(평가센터)기법. 한경비지니스. 2008년 4월 11일.

이영식, 신상근 (2004). 다변량 일반화가능도 이론에 의한 말하기 시험의 타당도와 신

뢰도에 관한 연구. Foreign Languages Education, 11, 249-265.

임대열 (2004). Assessment Center 프로그램의 개발 및 운영. 한국 산업 및 조직심리학회 추계학술대회 발표집(p.340).

조재윤 (2009). 일반화가능도 이론을 이용한 쓰기 평가의 오차원 분석 및 신뢰도 추정 연구. 국어교육, 128, 325-357.

지은림 (1999). 사회과 보고서 수행평가를 위한 총체적 채점과 분석적 채점의 비교. 교육평가연구, 12, 11-24.

지은림, 김성숙 (2005). 초등학교 수행평가의 교육적 효과와 활용 방식. 교육평가연구, 12, 173-191.

Arthur Jr., W., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical re-examination of the assessment center construct-related validity paradox. *Journal of Management, 26*, 813-835.

Borman, W. C. (1982). Validity of behavioral assessment for predicting recruiter performance. *Journal of Applied Psychology, 67*, 3-9.

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology, 91*, 1114-1124.

Bowler, M. C., & Woehr, D. J. (2008). *Evaluating assessment center construct-related validity via variance partitioning. In B. J. Hoffman (Chair), Reexamining Assessment Centers: Alternate Approaches.* Paper presented at the 23rd annual meeting of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Bowler, M. C., & Woehr, D. J. (2009). Assessment center construct related validity: Stepping beyond the MTMM matrix. *Journal of Vocational Behavior, 75*, 173-182.

Bray, D. W., Campbell, R. J. (1968). Selection of salesmen by means of an assessment center. *Journal of Applied Psychology, 52*, 36-41.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: American College Testing.

Brennan, R. L. (2001a). *Generalizability theory*. New York: Springer.

Byham, W. C. (1970). Assessment center for spotting future managers. *Harvard Business Review, 48*, 150-160, plus appendix.

Cascio, W. F. (2002). Changes in workers, work, and organizations. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Comprehensive handbook of psychology, 12 : Industrial and organizational psychology* (pp. 107-130). New York: Wiley.

Cohen, B. M., Moses, J. L., & Byham, W. C. (1974). *The validity of assessment centers: A literature review*. Monograph II. Pittsburgh, PA: Development Dimensions Press.

Crick, J. E. & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system.* ACT Technical Bulletin, 43, The American College Testing Program, IA.

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

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C.

(1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137-163.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 74,* 493-511.

Gibbons, A. M., Rupp, D. E., Baldwin, A., & Holub, S. A. (2005, April). *Developmental assessment center validation: Evidence for DACs as effective training interventions.* Paper presented at the 20th annual meeting of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Hardison, C. M., & Sackett, P. R. (2004). *Assessment center criterion-related validity: A meta-analysis update.* Paper presented at the 19th annual conference of the Society Industrial and Organizational Psychology, Chicago, Ill.

Howard, A. (1995). *The Changing nature of work.* San Francisco, CA: Jossey-Bass.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Jackson, D. J. R., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance, 18*, 213-241.

International Task Force on Assessment Center Guidelines (2009). Guidelines and Ethical Considerations for Assessment Center Operations. *International Journal of Selection and Assessment, 17*, 233-253.

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89,* 377 – .385.

Lievens, F. & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait -multimethod studies. *Journal of Applied Psychology, 86*, 1202-1222.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investingation of study characteristics. *Personnel Psychology, 37*, 407-422.

Thornton, C. G., III, Byham, W. C. (1982). *Assessment centers and managerial performance*. New York: Academic Press.

Thornton, G. C., III, & Rupp, D. E. (2003). Simulations and assessment centers. In J. C. Thomas (Ed.), & M. Hersen (Series Ed.), *Comprehensive handbook of psychological assessment, Vol. 4: Industrial and organizational assessment* (pp. 318-344). New York:

Thornton, G. C., III, & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Erlbaum.

1차 원고접수 : 2010. 3. 9
2차 원고접수 : 2010. 4. 27
최종게재결정 : 2010. 5. 23

한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2010. Vol. 23, No. 2, 225-249

Reliability and Validity of Nested-Designed Assessment Center

Chang-Goo Heo **Kang-Hyun Shin**

Ajou University

An Assessment Center(AC) consists of a standardized evaluation of behavior based on multiple evaluations including: job-related simulations, interviews and psychological tests. Recently, as many companies have been adopted AC as a selection tool, the importance of AC have been rapidly increased. The general frame work of AC used is called "nested model". That is, selecting a competence for assessment identifies exercises. However, there have been little researches conducted on the validity of nested AC. Therefore, this study is intended to the validity of nested AC using generalizability theory with G and D study and confirmatory factory analysis. In this study, we use two types of data: AC scores of applicants(n=1,249) and AC scores of incumbents(team manager=105, team member=200). The AC of this study is designed of Structured Interview(SI), which three competencies(responsibility, activity, innovation) are rated by three assessors and Presentation(PT) and Group Discussion(GD), which two competencies(information processing, problem solving / harmony, communication skill) are rated respectively by two assessors. In the result, generalizability analyses indicated that the reliabilities of three exercise were acceptable and exercise effects existed, on the other hand, dimension effects were found especially in SI. But, rater effect was almost not existed. Confirmatory factor analyses results were consistent with the generalizability results. That is, the dimensions of AC could be discriminated, and showed a differential predictive validity. Implications for future research and practice are discussed.

Key words : assessment center, nested-design, generalizability, d study, g study