한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2011. Vol. 24, No. 4, 785-807

대학선발장면에서 평가센터의 신뢰도 및 구성개념 타당도 분석

허 창 구†

아주대학교 사회과학연구소

본 연구는 대입선발을 목적으로 개발된 평가센터를 이용하여, 평가센터의 신뢰도와 타당도를 확인함으로써 대학장면에서의 평가센터 활용성을 검증하였다. 연구에 이용된 자료는 B 대학의 입학사정관제 모의전형에 응시한 지원자 60명에 대한 평가자 30명의 평가자료이다. 평가자 3인이 한 팀이 되어, 평가팀 별로 6명의 지원자에 대해 3가지 평가과제(구조화 면접, 프레젠테이션, 그룹토의)를 진행했다. 결정연구(D study)를 이용한 평가센터의 신뢰도는 양호하게 나타나 현재의 평가설계가 적절함을 보여주었다. 본 평가센터의 타당도 측면에서 재학생 대상 설문조사 결과 본 평가센터에서 평가하는 역량이 대학에서의 학업수행과 적응에 중요한 역량이라고 답하여 내용타당성을 보여주었으며, 평가센터 참가자 대상 설문조사 결과 평가의 공정성과 평가의 적절성 및 참여동기 측면에서 본 평가센터의 결과타당도의 일면을 보여주었다. 일반화 연구(G study)를 통해 각 평가과제에서 평가차원(21.1%)과 평가자(10.2%)의 영향을 비교한 결과 평가차원의 영향이 상대적으로 높게 나타났으며, MTMM 분석 결과는 MTHM(.51)은 5개 상관 모두 유의했고, HTMM(.55)은 8개 중 7개 상관이 유의했고, HTHM(.36)은 23개 중 8개 상관만이 유의하게 나타나, 평가센터에 평가차원 효과(dimension effect)와 평가과제 효과(exercise effect)가 모두 존재함을 알 수 있었다. 또한 평가센터 점수와 참가자의 여러 관련 변인들의 상관관계를 분석한 결과, 학업역량 중 자기주도적 학습은 지원자의 국어 및 영어 내신성적과 상관을 보여주었으며, 토론능력은 대중불안과 부적상관을 보여주었다. 인성역량인 도전정신은 지원자의 외향성, 자기효능감, 교외수상 횟수와 정적상관, 사회불안의 하위항목인 '낯선 것에 대한 두려움'과는 부적상관을 보여주었고, 창의성은 외향성과 정적상관, 적응력은 정서적 안정성과 정적상관을 보여주었다. 그룹토의에서 평가한 대인관계 역량은 수행불안 및 대인불안과 부적상관을 자기효능감과 정적상관을 보여주었다. 논의에서는 본 연구의 제한점과 향후연구 방향을 제시하였다.

주요어 : 평가센터, 입학사정관 전형, 일반화 가능도, D 연구, G 연구, MTMM

† 교신저자 : 허창구, 아주대학교 사회과학연구소, 경기도 수원시 영통구 원천동 산5번지, hck@ajou.ac.kr

최근에는 한국의 대입선발에도 새로운 바람이 불고 있다. 그 대표적인 예로 '입학사정관제'를 들 수 있는데, 입학사정관제도란 대학이 고등학교 교육과정 및 대학의 학생선발 방법 등에 대한 전문가를 채용하고, 이들을 활용하여 학생의 성적, 개인 환경, 잠재력 및 소질 등을 종합적으로 판단하여 신입생을 선발하는 제도이다(교육과학기술부, 2007). 한국의 입학사정관제도는 대학들의 성적위주 학생선발로 인해 발생한 입시위주의 교육흐름을 탈피하고, '발전 가능성이 큰 학생'을 선발하기 위해 학생의 다양한 잠재력을 파악하여 선발함으로써 입시위주의 교육흐름을 극복하기 위한 노력으로 도입되었다.

한국은 2008년부터 10개 대학에서 입학사정관제 전형으로 신입생을 선발하였으며, 2009년에는 41개 대학, 2010년 90개 대학, 2011년 118개 대학으로 확대되어가고 있으며, 각 대학은 잠재능력 우수자, 리더십, 자기추천서, 리더십우수자, 네오르네상스, 기회균형선발, 외국인 학생, 특수교육대상자 특별전형, 대안학교, 다빈치인재전형, 입학사정관제 전형 등 개별대학의 특성을 살려 다양한 전형제도를 마련하고 있다(한국대학교육협의회, 2008). 이들 대학이 이용하는 주요 전형방법으로는 서류(학생부, 자개소개서), 면접, 토의(분임토의, 집단토의), 인·적성검사 등이 있다.

입학사정관제도가 교과 성적보다는 학생 개개인의 적성, 소질, 잠재능력 등을 고려함으로써 교육 본래의 목적으로 방향을 전환했다는 점이 매우 고무적인 일임에도 불구하고, 한편으로는 과연 '어떻게' 잠재력 있는 인재를 선발하고 평가할 것인가? 라는 질문과 함께 많은 문제점들과 풀어야 할 과제들이 도출되었다. 이러한 문제점 중 가장 민감하고 중요한 문제로 지적되어온 것이 바로 입학사정관제도의 공정성과 객관성 및 신뢰성 확보의 문제라고 할 수 있으며, 이것은 본 제도가 정착하기 위한 가장 필수적인 조건이기도 하다(유현주, 2009).

본 연구는 입학사정관제 전형장면에 평가센터를 적용하여, 평가결과의 신뢰도와 타당도를 검증함으로 평가센터의 대입선발 장면으로의 적용 가능성을 확인하였다. 먼저, Binning과 Barrett(1989)이 말한 바와 같이 다각적인 측면에서 타당도 자료를 확보하였다. 평가센터의 내용타당도를 확인하기 위해 재학생을 대상으로 본 연구에서 측정한 평가차원이 성공적인 대학생활에 얼마나 중요한지를 조사하였고, 결과타당도 측면에서 평가센터 참가자를 대상으로 평가의 공정성과 평가의 적절성 그리고 그들의 참여동기를 조사했다. 또한 구성개념 타당도를 확인하기 위해 지원자의 평가차원 점수와 그들의 성격(five factor) 및 여러 인성검사 점수(우울, 심리적 자원)의 상관관계를 분석하여 평가차원의 차별적 평가 여부(구성개념 타당도)를 확인했으며, 일반화가능도 연구의 G 연구를 통해 각 평가과제에서 평가센터 점수에 미치는 평가차원의 영향정도를 확인하였다. 다음으로 일반화가능도 연구의 D 연구를 통해 평가센터의 신뢰도를 확인하고, 평가센터 설계의 적절성을 살펴보았다.

평가센터

인재선발 중요성이 커짐에 따라 "적합한 사람(Right People)"을 선발하기 위한 도구도 점차 발전해 왔는데, 그 중에서 평가센터 방식은 지원자들을 평가하고 개발하는 포괄적이고 탄력적인 도구로써, 복잡한 특성의 측정이

가능하고, 참가자들에 의해서도 공정하다고 인식되고 있으며(안면타당도), 편파효과(adverse impact)가 거의 없고, 다양한 준거(수행, 잠재력, 훈련성공, 승진)를 예측해주는 것으로 연구되었다(Gaugler, Rosenthal, Thornton, & Bentson, 1987).

평가센터는 구조화면접(structured interview)과 함께 역량을 가장 잘 평가해주는 도구로 여겨지고 있다. 일반적으로 역량 중 인성과 관련된 역량(예, 도전성, 창의성, 변화적응 등)은 경험기반 구조화 면접을 통해 평가하기 적합하며, 직무수행과 관련된 역량(예, 자료분석력, 문제해결력, 의사전달력 등)은 모의상황으로 평가하기 적합하다. 평가센터는 엄격한 의미에서 이러한 모의상황(simulation exercise)으로 구성된 평가방식이다. 한편, 넓은 의미로는 '역량모델링을 통해 도출된 역량'을 '다수의 훈련된 평가자'가 '모의활동이 포함된 다수의 평가기법'으로 평가하는 것으로 볼 수 있다(International Task Force on Assessment Center Guidelines, 2009). Thornton과 Rupp(2006)은 평가센터의 필수적인 구성요소로 직무분석을 통한 평가차원의 도출(job analysis), 다수의 평가기법(multiple assessment techniques), 모의활동(simulated exercises), 다수의 훈련된 평가자(multi trained assessors), 행동적 반응과 관찰(behavioral response and observations), 관찰내용의 통합(integration of observations)을 제시하고, 이러한 사항 이외에 평가센터를 대표하는 유일한 평가차원의 구성이나 평가과제의 구성은 존재하

표 1. 평가센터의 기본 구성요소

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 모의활동이 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

(참고) International Task Force on Assessment Center Guidelines(2009), 허창구, 신강현(2010)에서 재인용.

지 않는다고 말하고 있다. 본 연구에서 사용된 평가센터는 2009년에 발표된 평가센터 운영방식의 가이드라인(International Task Force on Assessment Center Guidelines, 2009)에 맞추어 광의의 평가센터 개념으로 설계되고 실행되었다.

평가센터를 대학교 장면에 적용한 대규모 연구는 1985년에 AACSB(American Assembly of Collegiate Schools of Business)에 의해 시작된 Skills Diagnostic Program(SDP)으로, 이 프로그램은 경영학 전공 학생들의 학업결과물을 평가하기 위한 것이었다(Riggio, Aguirre, & Mayes ,1997). 그 밖에 여러 대학에서 학생을 평가하는데 평가센터를 활용했는데 대부분 학업결과에 혹은 개발결과에 대한 평가를 평가센터로 실시한 경우이며(Bartels, Bommer, & Rubin, 2000; Kottke & Shultz, 1997; Loacker, 1991; Mullin, Shaffer, & Grelle, 1991; Riggio, Aguirre, & Mayes, 1997), 선발을 목적으로 평가센터를 적용한 사례는 찾아보기 힘들다.

한편, 한국대학교육협의회(2009)는 입학사정관제의 공정성 및 신뢰성 확보를 위해 가이드라인을 제안했는데, 이에 따르면, 입학사정관제는 사전공지, 서류심사, 심층면접/토론, 최종선발의 과정을 거치며, 평가요소로는 학생의 특성, 대학과의 적합성, 학생의 교육환경을 평가하도록 한다. 여기서 특히 학생의 특성은 인지적 특성(사고력, 적성 및 역량, 표현력), 정의적 특성(인성, 흥미, 태도), 잠재력, 미래성장가능성, 전공 적응 가능성 등으로 평가하도록 제안하고 있다. 또한, 이러한 전형의 실행원칙으로 '다수의 평가자에 의한 다단계 평가'를 적용하도록 권고하면서 대학별로 구체적인 계획을 수립・운영하도록 하고 있다. 이는 다수의 평가자, 다수의 평가역량, 다수의 평가도구를 기반으로 하는 평가센터의 방식에 매우 부합하며, 잠재능력이라 할 수 있는 역량의 평가에 평가센터 방식이 매우 타당한 도구임(Spencer & Spencer, 1993)을 감안할 때 평가센터를 입학사정관제 전형방식으로 적용하는 것이 가능하다고 할 수 있다.

평가센터의 신뢰도

지난 50년 동안 축적된 자료에 근거할 때 평가센터는 신뢰롭고 타당하다(Thornton & Rupp, 2006). 평가자들은 특정 역량과 관련된 해동을 관찰하고 평가하기 위해 훈련받으며, 이들의 평가는 신뢰도와 타당도를 보여주고 있다. 평가센터에 참여한 두 명의 평가자 간 평정합의(interrater agreement)의 평균은 .83에 이른다(Ladd, Atchley, Gniatczyk & Baumann, 2002). 특히 평가센터의 요소인 참조틀(Frame Of Reference, FOR) 교육은 역량평가의 신뢰도를 향상시켜, 커뮤니케이션(.76→.83), 의사결정(.69→.89), 리더십(.65→.75) 역량평가에서 신뢰도가 향상되는 것을 보여주었다(Schleicher, Day, Mayes, & Riggio, 2002).

평가센터는 다양한 모의활동을 통해 이루어지는 수행평가라고 할 수 있는데, 일반적으로 지원자가 지닌 다수의 역량(평가차원)을 다수의 평가과제를 통해 다수의 평가자가 평가하는 다각적인 평가가 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, 피평가자의 반응을 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. 더욱이 수행평가 결과를 선발과 같은 의사결정에 활용할 경우 수행평가 결과의 신뢰도 검증은 더욱 중요해진다(남명호, 2002).

평가센터의 타당도

평가센터는 선발되거나 승진되었을 때 성공적인 직무수행을 보일 것으로 생각되는 사람들의 전반적인 속성들을 측정한다. 이를 평가센터 전반점수(Overall Assessment Rating, OAR)라 하는데, 이는 세부적인 평가차원 점수들로 이루어져 있다. 이러한 평가차원들은 직무분석 혹은 역량모델링을 통해 도출되기 때문에 평가센터는 높은 내용타당도를 보여주고 있다(Thornton & Rupp, 2006). 준거관련 타당도 측면에서 OAR과 준거와의 상관관계는 직무수행과 .33~.64, 훈련수행과 .35~.48, 승진과 .40~.41의 범위로 나타났다(Borman, 1982; Bray & Campbell, 1968; Byham, 1970; Cohen, Moses, & Byhan, 1974; Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hunter & Hunter, 1984; Schmitt, Gooding, Noe, & Kirsch, 1984). 이에 Thornton과 Rupp(2006)은 평가센터가 관리자로서의 성공이나 수행에 대해 지니는 준거관련 타당도에는 의심의 여지가 없다고 결론지었다.

한편, 구성개념 타당도에 있어서는 여전히 논쟁이 진행되고 있는 중인데, 이는 평가센터의 점수가 평가센터에서 측정하는 평가차원들에 의해 나타나는 것인지, 아니면 평가센터에 포함된 평가과제에 의해 나타나는 것인지에 대한 것이다. 평가센터의 구성개념 타당도가 확보되기 위해서는 평가차원효과(dimension effect)가 평가과제효과(exercise effect)보다 우세해야 한다. 지금까지의 연구결과는 어떤 연구에서는 "평가과제 효과"가 두드러지게 나타났다(Lance, Foster, Thoresen, & Gentry, 2004; Lievens & Conway, 2001; Sackett & Tuzinski, 2001). 이들은 이러한 "평가과제 효과" 우월성을 근거로 평가센터에서 역량모델의 불필요성을 주장하지만, 이러한 주장은 "평가과제"에 기반한 평가센터의 타당성을 입증할 실증자료가 없다는 비판을 받아왔다(Thornton & Rupp, 2006). 하지만, 최근 평가과제에 기반한 평가센터의 실증연구가 시작되고 있다(Jackson, Stillman, & Englert, 2010). 한편, 다른 연구에서는 "평가과제 효과"와 "평가차원 효과"가 모두 나타났다(Donahue, Truxillo, Cornwell, & Gerrity, 1997; Lievens & Conway, 2001; Kudisch, Ladd, & Dobbins, 1997; Louiselle, 1980). 이들의 연구결과는 평가센터에서 평가과제의 특성에 의한 "평가과제 효과"가 물론 강하게 나타나긴 하지만, 역량평가에 의한 "평가차원 효과"또한 존재하므로 평가센터에서 평가역량은 여전히 필요함을 보여주었다. 더욱이 Gibbons, Rupp, Baldwin, 및 Holub(2005)의 연구에서는 평가차원을 이루는 하위 평가차원들이 차별적으로 평가되고 있음을 보여줌으로써 "평가차원 효과"의 증거를 보여주었다. 따라서 평가센터의 구성개념 타당도 입증은 평가센터에서 역량모델링의 필요성 및 역량의 차별적 평가의 필요성을 확인하는 의미있는 절차라고 할 수 있다.

그 밖에 다양한 평가도구에 대한 참가자의 반응은 결과타당도 측면에서 중요한데 그 이유는 이러한 반응이 조직에 대한 태도로 이어질 것이기 때문이다(Bauer, Maertz, Dolen, & Campion, 1998; Howard, 1997).

연구방법

평가센터 도구 개발

역량모델링

평가센터 개발에 앞서 대학이 선발하고자 하는 인재상을 첫째, 입학 후 학업을 잘 수행하고, 둘째, 환경변화에 잘 적응할 수 있는 사람으로 요약하고 이와 관련된 핵심적 역량을 도출하기 위해 FGI(Focus Group Interview)를 실시하였다.

먼저, 학업관련 특성을 도출하기 위해 우수/비우수한 학업성취를 보인 학생들의 차별적 행동특징을 알아보았다. 이를 위해 2008학년도 수시전형으로 입학 후 1년 동안 우수(4.0이상)/비우수(2.0이하)한 학업수행을 보인 두 집단의 학생들을 대상으로 인터뷰를 실시했다. 다음으로 적응관련 특성을 도출하기 위해 상담전문가 및 상담수련생(대학원생)을 대상으로 인터뷰를 실시했다. 인터뷰 내용의 분석을 통해 학업과 적응 측면에서 우수 집단과 비우수 집단을 구별해 주는 주요한 특성 리스트를 산출했다.

인터뷰 결과로 도출된 특성들을 바탕으로 국내외 기업 및 대학의 역량관련 문헌을 바탕으로 우수한 대학인재가 갖추어야 할 역량을 도출했다. 핵심역량 선정 과정에서 연구자와 동료 연구자, 대학 입학관계자, 및 외부 전문가 등이 협의를 거쳤다. 협의과정에서 인터뷰를 통하여 추출된 학업 우수자의 행동특성 및 적응 우수자의 행동특성, 직업사회에서 요구하는 주요 핵심역량, 대학 생활에서의 주요활동 등이 고려되었다.

최종으로 도출된 역량모델은 총 4개의 역량군(학업, 개인, 대인, 적응)이며, 각 역량군에는 총 10개의 역량이 포함되었다. 구체적으로 '학업역량군'은 학업을 잘 수행할 사람을 선별하기 위한 역량으로 자기주도적 학습, 분석적 사고, 발표력, 토론능력을 평가한다. '대인역량군'은 대인관계가 원만할 사람을 선별하기 위한 역량으로 타인배려적 의사표현, 적극적 경청, 의견조율을 평가한다. '개인역량군'은 창의적이고 도전적인 사람을 선별하기 위한 역량으로 창의성과 도전정신을 평가한다. '적응역량군'은 환경변화에 대한 적응력이 높은 사람을 평가하기 위한 역량으로 환경변화적응을 평가한다. 또한 개념이 포괄적인 몇몇 역량은 세부역량으로 구분되었는데, '자기주도적 학습'은 3가지 세부역량(학습계획설정, 지속적 실행, 자기점검)으로, '분석적 사고'는 2가지 세부역량(자료이해, 통합적 사고), '발표력'은 3가지 세부역량(효과적 의사전달, 발표자료의 구성, 바른자세)로 세분화 되었다.

도출된 역량의 내용타당도를 확인하기 위해 재학생을 대상으로 성공적인 대학생활을 위한 각 평가차원의 중요도를 1점(중요하지 않다)에서 5점(매우 중요하다)으로 물은 결과, 자기주도적 학습의 중요도 점수가 4.5점으로 가장 높았으며, 적극적 경청 4.4점, 환경변화 적응력이 4.3점, 분석적 사고와 도전정신, 적극적 경청, 타인배려적 의사표현, 의견조율이 4.2점, 발표력과 토론능력이 4.1점의 순이었다. 결과적으로 모든 평가역량의 중요도 평균점수가 4점(중요하다) 이상으로 나타나, 본 연구에서 평가하는 역량이 학업수행에 중요한 역량들임을 알 수 있었다.

평가과제 개발

평가과제로는 총 3가지가 개발되었다(구조화 면접, 프레젠테이션, 집단토의).

먼저 구조화 면접(Structured Interview, SI)은 피평가자가 지닌 다양한 잠재역량을 심층적으로 평가하기 위한 도구이다. 구조화 면접 질문세트는 해당 역량 별로 피평가자의 경험을 도출해 낼 수 있는 리드질문(Lead Question)과 리드질문에 대한 피평가자의 답변에 대해 심층적인 사항을 물을 수 있는 탐침질문(Probing Question)으로 구성하였다. 각 탐침질문들은 해당 역량에 대한 행동기준 평정척도(BARS)에 근거해 평가를 할 수 있는 피평가자의 구체적인 경험을 유도하게끔 고안하였다.

다음으로 프레젠테이션(Presentation, PT)은 피평가자들이 대학생활에서 수시로 접하게 될 발표상황에서의 수행을 예측하기 위한 도구이다. 피평가자들은 주어진 자료를 바탕으로 자신의 해결안을 도출하여 프레젠테이션 하도록 했다.

마지막으로 그룹토의(Group Discussion, GD) 또한 대학생활에서 자주 접하게 될 상황으로 채택되어 개발되었다. 그룹토의의 주제는 참가자들이 이견을 내세워 토론할 수 있는 주제였으며, 토론의 활성화를 위해 참가자들은 자신의 주장을 펼치면서도 합의안을 도출하도록 요구받았다.

평가기준 개발

평가기준은 행동기준평정척도(Behaviorally Anchored Rating Scale: BARS) 형식으로 제작하였다. BARS는 Smith와 Kendall(1963)이 산업장면에서의 행동을 측정하는 데 처음으로 사용된 평정척도의 한 종류로서, 척도상의 눈금이나 눈금간의 위치에 부합하는 행동을 하나의 문장으로 만들어서 그 위치에 기재하도록 하고, 그 위치의 의미를 부여하는 방식이다. 추상적인 용어나 특성(매우 우수, 우수, 보통 등)을 사용하기보다, 누구나 이해할 수 있는 기준을 제시해준다는 점에서 유용하다.

BARS는 세부역량 별로 제작되었으며, 5점 평가 척도 중, 1점-3점-5점에 대한 행동기준을 제작하고, 그에 따른 행동 예시문들을 제작하였다. 행동예시문은 BARS를 근거로 실제 평가장면에서 발견될 수 있는 행동들로써 평가자의 정확한 평가에 도움을 줄 수 있다.

파일럿 테스트

개발한 도구의 현장 적용 가능성 점검 및 개선사항을 파악하기 위해 각 도구 별로 구조화 면접 10회, 프레젠테이션은 10회, 그룹토의는 4회에 걸친 파일럿 테스트를 실시했다. 각 테스트는 모의 지원자를 대상으로 실제 전형과 동일하게 진행되었으며, 면접 장면을 녹화하여 개선사항 점검에 활용했다.

평가자 교육

평가센터를 운영하기 위해서는 훈련된 평가자가 필수적이다. 평가센터는 다수의 피평가자가 지닌 다수의 역량을 다수의 평가자가 다수의 도구를 이용해 평가하는 방식이기 때문에 평가자간의 평가기준의 균일성이 요구된다. 이를 '참조틀'이라하는데 평가자 교육을 통해 참조틀을 통일하는 것이 1차적 목적이며, 평가역량 및 평가과정에 대한 숙지가 2차적 목적이다.

평가자 교육은 현직 교수 30명을 대상으로 이틀 간 진행되었다. 교육은 5개의 실습팀으로 구성하여 이론교육 및 모의지원자를 대상으로 한 구조화 면접, 프레젠테이션 및 그룹토의 실습을 실시했다. 평가자 교육 결과 후 모의평가 실습에서 평가자 간 합치도(cronbach-α)는 평균 0.88로 높게 나타났다.

연구 대상자

본 연구에 이용된 자료는 B 대학의 입학사정관제 모의전형에 참가한 60명(남30, 여30)에 대한 평가자(교수) 30명의 평가자료이다. 실제 분석에는 평가자의 기록이 불완전한 자료를 제외한 54명의 자료가 이용되었다. 참가자들은 고 2에 재학 중인 학생들이었으며, 지원자들 중에서 입학사정관들에 이해 선정되었다. 평가자들은 인문, 사회 및 이공에 속한 각 학과의 교수 중에서 본 평가과정에 호의적인 교수들이 참여하였다. 평가자 3인이 한 조가 되어 총 10개의 평가팀이 구성되었으며, 평가팀 별로 6명의 지원자에 대해 3가지 평가과제(구조화 면접, 프레젠테이션, 그룹토의)를 진행했다.

분석자료

구조화 면접(SI)에서는 평가자 3인이 지원자 1인을 대상으로 30분간 심층면접을 거치며 4가지 평가차원(자기주도적 학습, 도전정신, 창의성, 환경변화 적응)을 평가했다. 프레젠테이션(PT)에서는 지원자가 40분간 주어진 자료를 분석한 후 5분 발표와 10분 질의응답을 실시했고, 그 과정에서 2가지 평가차원(분석적 사고, 발표력)을 평가했다. 그룹토의(GD)에서는 6명의 지원자가 주어진 토론주제에 대해 10분간 검토시간을 가진 후, 30분 간 리더 없는 그룹토의(leaderless group discussion)을 진행하였고, 그 과정에서 4가지 평가차원(토론능력, 적극적 경청, 의견조율, 타인배려적 의사표현)을 평가했다.

모든 평가는 BARS 형태로 준비된 평가기준에 따라 절대평가로 이루어졌으며, 개별 평가자의 평가 후에는 피평가자의 세부 활동 점수에 대한 '평가자 회의'를 진행했다. 이때 평가자 회의는 정보교환을 목적으로 하며 평가점수 합의는 요구되지 않았다. 본 연구는 평가

표 2. 평가센터 역량 및 측정방법

역량군	역량	측정방법		
		SI	PT	GD
학업	1. 자기주도적 학습	X		
	2. 분석적 사고		X	
	3. 발표력		X	
	4. 토론능력			X
개인	5. 도전정신	X		
	6. 창의성	X		
대인	7. 적극적 경청			X
	8. 의견조율			X
	9. 타인배려적 의사표현			X
적응	10. 환경변화적응	X		

SI(Sturctured Interview), PT(Presentation), GD(Group Discussion)

X 는 특정 평가과제에서 측정되는 역량임

표 3. 평가센터 참가자 대상 사후설문 검사내용

검사명	출처	세부항목
성격 5요인	Donnellan, Oswald, Baird & Lucas(2006)	외향성, 개방성. 성실성, 호감성, 정서적 안정성
한국판 아동청소년용 사회불안척도	문혜신 · 오경자(2002)	수행불안, 회피행동 및 사고(대중불안), 낯선 것 두려움, 부정적 평가 두려움, 비주장성
심리적 자원 (Psycap)	Luthans, Avolio, Avey & Norman(2006)	자기 효능감, 희망, 탄력성, 낙천주의

자 회의 이전에 각 평가자가 산출한 세부 활동 점수를 이용하여 분석하였다.

또한, 본 평가센터 실시 약 한 달 후, 참여했던 60명에게 우편을 통해 설문조사를 실시했다. 전체 참가자 중 48명으로부터 응답이 회수되었다(회수율 66.7%). 설문조사의 내용은 평가과정의 공정성, 평가의 적절성, 평가참여 동기 등에 대한 질문이었다. 이는 본 평가센터의 결과타당도(consequential validity)를 알아보기 위한 것이었다. 결과타당도란 검사과정이 의도된 긍정적 결과를 낳는지 혹은 의도하지 않은 부정적 결과를 낳는지에 대한 것으로 평가센터의 경우 결과타당도가 높은 것으로 알려져 있다(Thornton & Rupp, 2006). 그 밖에 성격 5요인(five factor), 사회불안척도(social anxiety), 심리적 자원(psycap)을 측정했으며, 이 중 성격 5요인과 시리적자원은 해외 검사를 번안하여 사용하였다. 이들 변인은 각 역량 평가점수와의 관련성을 확인하기 위한 것으로 성격 5요인은 도전정신, 창의성, 적응력 등과의 관련성을, 사회불안척도는 적응력, 토론능력, 대인관계 등과의 관련성을, 심리적 자원은 도전정신, 적응력 등과의 관련성을 알아보기 위한 것이었다.

한편, 피평가자들이 제출한 서류심사 자료 중 내신등급과 교외수상실적을 이용해 '자기주도적 학습' 및 '도전정신' 평가점수의 수렴타당성을 확인하기 위해 분석에 이용하였다.

일반화가능도

평가센터와 같은 수행평가 점수에는 평가자, 평가과제, 평가차원 및 이들의 상호작용과 같은 다양한 요인이 영향을 미친다. 따라서 수행평가의 신뢰도는 위와 같은 요소들의 영향을 고려하여 산출할 필요가 있다. 이와 같은 수행평가의 다양한 영향요인에 대한 고려는 전통적인 검사이론에서는 수용될 수 없는 것으로 새로운 이론적인 접근이 도입되어야만 하는데, Cronbach, Rajaratnam와 Gleser(1963)는 다양한 요인이 영향 미치는 수행평가의 신뢰도를 산출하기 위한 일반화가능도 이론을 고안했다. 이러한 일반화가능도 분석은 크게 두 가지로 구분된다(일반화연구, 결정연구)

일반화 연구(G 연구)

일반화 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 분산분석을 실시한 결

과를 토대로 전체 평가점수에 영향을 미치는 각 국면의 분산크기를 추정하는 절차이다(남명호, 1996). 즉, 일반화 연구는 각 요인(facet)이 전체 점수에 미치는 영향정도를 추정해 주며 도구의 개발단계에서 많이 이용된다.

결정 연구(D 연구)

결정 연구는 G연구에서 추정된 각 국면의 변산성을 이용해 전체 도구의 신뢰도를 계산하고, 적절한 신뢰도 확보를 위해 갖춰야 할 각 요인(facet)의 수준(수)을 결정하는 연구로, 향후의 평가도구설계를 위한 제안점을 제공해 준다. 즉, 연구목적에 따라 선택된 오차원(예, 평가자, 문항, 과제)의 영향력에 따라 각 오차원의 수를 다르게 조절함으로써 알맞은 신뢰도 계수를 확보하기 위한 측정 조건을 결정할 수 있다(허창구, 신강현, 2010). 결정 연구 결과로 제시되는 신뢰도 계수는 상대평가 상황의 경우 일반화가능도 계수를 사용하며 전통적인 신뢰도 계수(cronbach α)와 동일하게 정의되기 때문에(Cronbach et al, 1972) .7~.8 정도를 적절한 신뢰도 수준으로 간주한다. 한편 절대평가 상황에서는 의존도 계수를 사용하는데, 상대오차 보다 큰 절대오차를 사용하기 때문에 일반화가능도 계수보다 낮게 산출되며, 따라서 .6~.8 정도를 적절한 신뢰도 수준으로 간주한다(조재윤, 2009).

연구결과

신뢰도 분석

평가자 간 합치도(Cronbach-α)

평가차원 별 평가자 내적 합치도는 그룹토의(GD)에 속한 한 개의 평가차원(타인배려적 의사표현)에서 .59 로 다소 낮은 합치도를 보여주었으나, 그 밖의 평가차원에서는 .73~.91로 높은 평가자 내적 합치도를 보여주었으며, 평가과제 별 분석에서도 .89~.93로 높은 평가자 내적 합치도를 보여주었다.

평가과제 별 신뢰도(D Study: p×R×D)

구조화 면접의 신뢰도. 본 연구의 구조화 면접에서는 3명의 평가자(R)가 6개 세부 평가차원(D)을 평가하도록 설계되어 있다. 수행평가 방식으로 이루어진 본 자료에서 효율적인 국면의 수준을 결정하기 위한 p×R×D 결정연구는 평가자의 수를 2명부터 4명까지, 평가차원의 수를 2개에서 6개까지 조절하여 설계하였다. 신뢰도 계수로는 본 자료가 절대평가 자료이므로 의존도 계수를 확인하였다. 그 결과 의존도 계수는 .719로 양호하게 나타나, 신뢰도 향상을 위해 평가자나 평가차원의 수를 증가시킬 필요성이 없는 것으로 나타났다. 평가자수를 현재와 같이 3명으로 유지할 경우 평가차원의 수가 3개 이상일 경우 의존도 계수가 기준인 .60 수준을 만족시켰으며, 평가차원의 수를 현재와 같이 6개로 유지할 경우 평가자의 수가 2명으로 줄어도 의존도 계수는 .673으로 기준을 만족시키는 것으로 나타났다.

프레젠테이션의 신뢰도. 본 연구의 프레젠테이션은 3명의 평가자(R)가 5개 세부 평가차원(D)을 평가하도록 설계되어 있다. 결정연구는 평가자의 수를 2명부터 4명까지, 평가차원의 수를 2개에서 6개까지 조절하여 설계하였다. 그 결과 의존도 계수는 .854으로 양호하게 나타났다. 한편, 평가자수를 현재와 같이 3명으로 유지할 경우 평가차원의 수가 2개까지

표 4. 평가자 간 평가점수 일치도(cronbach-α)

과제	평가차원	평가차원 별	과제별
구조화면접 (SI)	자기주도적 학습(계획, 실행, 점검)	.91	0.92
	도전정신	.86	
	창의성	.91	
	환경변화적응	.85	
프레젠테이션 (PT)	분석적 사고(자료이해, 통합적사고)	.91	0.93
	발표력(의사전달, 자료구성, 자세)	.90	
그룹토의 (GD)	토론능력	.80	0.89
	적극적 경청	.73	
	타인배려 표현	.59	
	의견조율 노력	.87	

감소해도 의존도 계수가 .769로 양호한 수준을 유지했으며, 평가차원의 수를 현재와 같이 6개로 유지할 경우 평가자의 수가 2명으로 줄어도 의존도 계수는 .825으로 매우 양호한 신뢰도를 보여주었다.

표 5. 국면의 조정에 따른 평가과제 별 신뢰도의 변화추이

평가자 수	평가 차원 수	SI	PT	GD
2	2	0.486	0.721	0.601
2	3	0.564	0.770	0.676
2	4	0.614	0.796	0.722
2	5	0.648	0.813	0.752
2	6	0.673	0.825	0.774
3	2	0.526	0.769	0.653
3	3	0.608	0.814	0.726
3	4	0.659	0.839	0.769
3	5	0.694	0.854	0.797
3	6	0.719	0.865	0.817
4	2	0.548	0.795	0.683
4	3	0.632	0.838	0.753
4	4	0.684	0.862	0.794
4	5	0.720	0.876	0.821
4	6	0.745	0.887	0.840

그룹토의의 신뢰도. 그룹토의에서는 3명의 평가자(R)가 4개 세부 평가차원(D)을 평가했다. 결정연구는 평가자의 수를 2명부터 4명까지, 평가차원의 수를 2개에서 6개까지 조절하여 설계하였다. 그 결과 의존도 계수는 .769으로 양호하게 나타났다. 평가자수를 현재와 같이 3명으로 유지할 경우 평가차원의 수가 2개까지 감소해도 의존도 계수가 .653로 신뢰로운 기준을 만족시켰으며, 평가차원의 수를 현재와 같이 4개로 유지할 경우 평가자의 수가 2명으로 줄어도 의존도 계수는 .722로 양호한 수준을 유지했다.

평가과제 전체의 신뢰도(D Study: p×R×E design). 효과적인 평가센터 설계를 위해 평가자 국면과 평가과제 국면을 대상으로 평가센터의 신뢰도를 추정하고 평가자 수와 평가차원 수의 변화에 따른 신뢰도 변화를 확인한 결정연구에서, 3명의 평가자가 3개의 과제를 평가하도록 되어있는 본 연구의 의존도 계수는 .650로 신뢰로운 수준이었다. 현재와 같이 평가자의 수를 3명으로 유지할 경우 평가과제의 수가 증가해도 의존도 계수에 미치는 영향은 매우 적었다. 한편, 평가 과제의 수를 3개로 유지할 경우 평가자의 수가 2명으로 감소하면 의존도 계수가 .554로 신뢰로운 기준에 미치지 못하는 것으로 나타났다. 즉, 평가자 수와 평가과제의 수를 정할 때, 신뢰도에 보다 큰 영향을 미치는 것은 평가자의 수였으며 평가자수는 3명이상으로 유지해야 .6이상의 신뢰로운 의존도 계수를 얻을 수 있었다.

표 6. 국면의 조정에 따른 평가센터 신뢰도 변화추이

평가자 수	평가 과제 수	AC
2	2	0.526
2	3	0.554
2	4	0.570
2	5	0.579
2	6	0.586
3	2	0.624
3	3	0.650
3	4	0.664
3	5	0.673
3	6	0.679
4	2	0.687
4	3	0.712
4	4	0.725
4	5	0.732
4	6	0.738

타당도 분석

피평가자의 반응

평가의 공정성을 묻는 질문에서 피평가자들은 평가과정에서 자신의 역량을 보여줄 기회를 부여받았다(4.08)고 답했으며, 평가과정이 모든 피평가자들에게 공정하게 진행되었다고 답했다(3.96점). 평가의 적절성을 묻는 질문에서도 이러한 평가과정에서의 우수한 수행이 대학생활에서의 우수한 수행으로 이어질 것이라고 답했으며(4.10점), 이러한 평가과정이 대입신입생 선발 방법으로 적합하다고 응답했다(4.31점). 이러한 결과는 피평가자들이 평가과정에 대해 절차공정성을 지각하고 있는 것으로 판단하는 근거가 될 수 있을 것이다.

구조화 면접(SI)의 일반화 연구

구조화 면접(SI)에서 평가점수에 영향을 미치는 평가자 국면(r)과 평가차원 국면(d)의 영

표 7. 평가과정에 대한 피평가자의 반응 (N=48)

	문항	평균	SD
문1	이번 평가과정은 지원자들이 자신의 역량을 보여줄 기회를 제공해 주었다. (평가의 공정성)	4.08	0.87
문2	이번 평가 과정은 모든 지원자들에게 동일한 방식으로 진행되었다. (평가의 공정성)	3.96	0.92
문3	이러한 평가에서 우수한 성과를 거둔 사람은 대학생활에서도 우수한 성과를 거둘 것 같다. (평가의 적절성)	4.10	0.91
문4	이러한 평가가 대학 신입생 선발을 위한 좋은 방법이라고 생각한다. (평가의 적절성)	4.31	0.62
문5	이번 평가를 수행하면서 좋은 성과를 거두려고 노력했다. (평가참여 동기)	4.77	0.47

향을 확인하기 위해, p×r×d 설계에 따른 G 연구를 실시했다. 평가자 국면을 살펴보면, 평가자의 주효과(r)는 전혀 나타나지 않았다. 그러나 피평가자와 평가자의 상호작용(p×r)은 11.0%로 나타났는데, 이는 평가자 자체에 따른 평가의 편향은 없으나 피평가자에 따라 평가자의 평가가 어느 정도 다르게 나타난다는 것을 말한다.

다음으로 평가차원 국면에서는 주효과(d)가 6.1%로 나타났으며, 피평가자와 평가차원의 상호작용 효과(p×d)가 26.2%로 높게 나타났다. 이 두 효과를 합하면 평가차원과 관련된 분산 추정치가 32.3%로서, 구조화 면접에서 피평가자들의 역량이 차별적으로 평가되었음을 의미한다. 한편, 평가자 국면과 평가차원 국면의 상호작용(r×d)은 0.3%로 거의 나타나지 않았다.

프레젠테이션(PT)의 일반화 연구

먼저 피평가자 국면을 살펴보면 49.2%로 높게 나타났다. 이는 피평가자의 전반적인 특성이 프레젠테이션 점수에 큰 영향을 미친다는 것을 의미한다. 평가자 국면을 살펴보면, 평가자의 주효과(r)는 전혀 나타나지 않았으나, 피평가자와 평가자의 상호작용(p×r)은 12.4%로 어느 정도 존재하는 것으로 나타났다. 이는 평가자 자체에 따른 평가의 편향은 없으나 피평가자에 따라 평가자의 평가가 어느 정도 다르게 나타난다는 것을 말한다.

다음으로 평가차원 국면에서는 주효과(d)가 1.9%로 매우 약하게 나타났으나, 피평가자와 평가차원의 상호작용(p×d)은 10.9%로 나타났다. 이것은 피평가자의 역량이 어느 정도 차별적으로 평가된다고 볼 수 있으나, 피평가자와 평가자의 상호작용 보다는 낮은 영향력이었다. 한편, 평가자 국면과 평가차원 국면의 상호작용(r×d)은 0.1%로 거의 없는 것으로 나타났다.

그룹토의(GD)의 일반화 연구

평가자 국면을 살펴보면, 평가자의 주효과(r)가 0.9%로 거의 나타나지 않았으며, 피평가자와 평가자의 상호작용(p×r)은 6.4%로 약하게

표 8. 평가과제 별 일반화 연구 분산성분 비율(%) 요약

국면	SI	PT	GD	평균
피평가자(p)	27.4	49.2	34.3	40.0
평가자(r)	0.0	0.0	0.9	0.3
평가차원(d)	6.1	1.9	0.0	2.7
피평가자×평가자(pr)	11.0	12.4	6.4	9.9
피평가자×평가차원(pd)	26.2	10.9	18.1	18.4
평가자×평가차원(rd)	0.3	0.1	1.8	0.7

존재하는 것으로 나타났다. 따라서 평가자 자체에 따른 평가의 편향은 없으나 피평가자에 따라 평가자의 평가가 약하게나마 다르게 나타난다는 것을 알 수 있다.

다음으로 평가차원 국면에서는 주효과(d)가 전혀 나타나지 않았으나, 피평가자와 평가차원의 상호작용 효과(p×d)는 18.1%로 어느 정도 영향력이 있는 것으로 나타났다. 이는 그룹토의가 피평가자의 역량을 차별적으로 평가해주고 있음을 말하는 것이다. 한편, 평가자 국면과 평가차원 국면의 상호작용(r×d)은 1.8%로 매우 약하게 나타났다.

평가차원 간 상관관계 분석

본 평가센터에서는 학업 역량군을 구성하는 자기주도적 학습, 분석적 사고, 발표력, 토론능력이 세 가지 평가과제에 걸쳐 평가되고 있어 구성타당도의 측면을 보여주는 MTHM (monotrait heteromethod)을 살펴볼 수 있으므로, 다특성다측정(MTMM) 상관관계 분석을 통해 평가차원의 수렴 및 변별타당도를 확인하였다.

다른 평가과제에서 측정된 동일 역량군에 속하는 평가차원들 간의 상관관계(MTHM)는 모두 유효했으며 상관관계의 크기는 평균 .509로 나타났다. 이는 동일한 역량군에 속한 평가차원 간의 수렴관계를 보여주는 것으로 평가센터를 구성하는 평가차원의 구성타당도를 반영하는 것이다. 다음으로 동일한 평가과제에서 측정된 상이한 역량군에 속한 평가차원 간의 상관관계(HTMM)는 8개 중 7개가 유효했으며 유효상관들의 평균은 .552로 나타났다. 자기주도 학습력과 창의성의 경우에는 동일한 평가과제(SI)에서 평가됨에도 불구하고 의미있는 상관을 보여주지 못했지만, 다른 상관관계에서 높은 값을 보여준 것은 평가과제의 효과 때문으로 볼 수 있다. 한편, 상이한 평가과제에서 측정된 상이한 역량군에 속한 평가차원 간의 상관관계(HTHM)는 23개 중 8개에만

표 9. 동일 역량군-다른 평가과제의 상관관계(MTHM)

(N=53)

MTHM 관계	상관관계
자기주도 학습(SI) - (PT)분석적 사고	.525**
자기주도 학습(SI) - (PT)발표력	.465**
자기주도 학습(SI) - (GD)토론능력	.440**
분석적 사고(PT) - (GD)토론능력	.491**
발표력(PT) - (GD)토론능력	.623**
유효 상관관계 수/해당 상관관계 수	5/5
유효 상관관계 평균	.509

표 10. 다른 역량군-동일 평가과제의 상관관계(HTMM) (N=53)

HTMM 관계			상관관계
자기주도 학습(SI)	-	(SI)도전정신	.414**
자기주도 학습(SI)	-	(SI)창의성	
자기주도 학습(SI)	-	(SI)변화적응	.451**
도전정신(SI)	-	(SI)변화적응	.624**
창의성(SI)	-	(SI)변화적응	.485**
토론 능력(GD)	-	(GD)타인배려적 의사표현	.690**
토론 능력(GD)	-	(GD)적극적 경청	.485**
토론 능력(GD)	-	(GD)의견조율	.610**
유효 상관관계 수/해당 상관관계 수			7/8
유효 상관관계 평균			0.552

표 11. 다른 역량군-다른 평가과제의 상관관계(HTHM) (N=53)

HTHM 관계			상관관계
자기주도 학습(SI)	-	(GD)적극적 경청	.024
자기주도 학습(SI)	-	(GD)타인배려적 의사표현	.365**
자기주도 학습(SI)	-	(GD)의견조율	.094
도전정신(SI)	-	(PT)분석적 사고	.154
도전정신(SI)	-	(PT)발표력	.267
도전정신(SI)	-	(GD)토론능력	.352**
도전정신(SI)	-	(GD)적극적 경청	.266
도전정신(SI)	-	(GD)타인배려적 의사표현	.397**
도전정신(SI)	-	(GD)의견조율	.296*
창의성(SI)	-	(PT)분석적 사고	.076
창의성(SI)	-	(PT)발표력	.154
창의성(SI)	-	(GD)토론능력	.129
창의성(SI)	-	(GD)적극적 경청	.015
창의성(SI)	-	(GD)타인배려적 의사표현	.149
창의성(SI)	-	(GD)의견조율	.135
변화적응(SI)	-	(PT)분석적 사고	.244
변화적응(SI)	-	(PT)발표력	.345*
변화적응(SI)	-	(GD)적극적 경청	.201
변화적응(SI)	-	(GD)타인배려적 의사표현	.392**
변화적응(SI)	-	(GD)의견조율	.179
분석적 사고(PT)	-	(GD)타인배려적 의사표현	.410**
분석적 사고(PT)	-	(GD)적극적 경청	.286*
분석적 사고(PT)	-	(GD)의견조율	.134
유효상관 수			8/23
유효상관 평균			.355

표 12. 평가차원과 관련변인의 상관관계 (N=53)

관련변인		SI						PT		GD	
		자기주도 학습			도전정신	창의성	적응력	분석적 사고	발표력	토론	대인
		계획	실행	점검							
성격	외향성	-.008	.096	-.030	**.326***	**.376***	.241	-.015	.147	.129	.282
	정서적 안정성	-.010	.010	.170	.215	.127	**.310***	.128	.189	.100	.117
사회불안	수행불안	.017	-.062	.021	-.156	.004	-.050	-.152	-.250	-.255	**-.375***
	대중불안	-.080	-.021	.031	-.181	-.077	-.245	.006	-.098	**-.301***	**-.302***
	낯선것두려움	.079	.078	.122	**-.387****	-.192	-.225	.055	-.078	-.141	-.268
심리자원	자기효능감	-.075	-.125	-.199	**.441****	.243	.208	-.061	.100	.227	**.311***
내신등급	국어	**.370***	.261	**.378***	-.033	-.123	**.348***	.205	.199	.219	.134
	영어	**.370***	.126	.222	-.112	-.169	.283	.171	.130	.148	-.054
	수학	.233	-.015	.086	-.223	-.225	.093	-.020	-.079	-.085	-.040
교외수상 횟수		-.032	.052	.091	**.364***	.240	.101	-.040	-.075	-.042	.073

유효했으며 유효상관들의 평균은 .355로 나타났다. 이 상관관계(HTHM)는 평가차원 효과와 평가과제 효과 중 어느 것과도 관련 없는 것으로써, 이 상관관계가 높을수록 평가차원의 구성타당도가 약하다고 할 수 있기 때문에, 본 연구결과와 같이 유효하지 않은 상관관계가 많이 나타나고, 유효 상관관계의 값도 MTHM 보다 낮게 나타난 것은 평가차원의 구성타당도를 지지해주는 증거라 할 수 있다.

평가차원과 관련변인의 상관관계 분석

평가차원 점수와 여러 관련변인의 상관관계를 통해 각 평가차원들의 차별성을 확인했다. 구조화 면접에서 평가한 자기주도적 학습의 경우 세 가지 세부활동(계획, 실행, 점검)으로 분리하여 평정하였기에 상관관계도 개별적으로 분석하였다. 그룹토의에서는 토론 진행 후 평가자들이 크게 두 가지 즉, 토론능력과 대인관계 역량(적극적 경청, 타인배려적 의사표현, 의견조율)을 기준으로 평가를 진행한 관계로 상관관계도 두 항목(토론, 대인관계)으로 묶어 분석하였다.

먼저, 구조화 면접에서 평가한 평가차원들을 살펴보면, 학업관련 평가차원인 자기주도적 학습(계획수립, 지속적 실행, 자기점검)은 내신등급과 관련성을 보여주었다. 세부역량별로 '계획수립'은 국어 및 영어 내신등급과 상관관계(각각 r=.370, p<.05)를 보여주었으며, '자기점검'은 국어 내신등급과 상관관계(r=.378, p< .05)를 보여주었다. 즉, 학습계획 수립과 자기점검에서 높은 역량점수를 받은 참가자는 국어내신과 영어내신 등급이 우수했다. 다음 역량인 '도전정신'은 외향성(r=.326, p<.05), 사회불안의 낯선 것 두려움(-.387, p<.05), 심리적 자원의 자기효능감(r=.441, p<.01), 교외 수상실적(r=.364, p<.05)과 상관

관계를 보여주었다. 즉, 도전정신에서 높은 역량점수를 받은 참가자는 외향성이 높았고, 낯선 것에 대한 두려움이 적었으며, 자기효능감이 높았고, 교외 수상실적이 많았다. 다음 역량인 '창의성'은 외향성(.376, $p<.05$)과 정적 상관관계를 보여주었다. 다음 역량인 '변화적응'은 정서적 안정성($r=.310$, $p<.05$)과 정적인 상관관계를 보여주었다. 특이한 점은 '변화적응'과 국어내신 등급이 상관관계($r=.348$, $p<.05$)를 보여주고 있다는 점이었는데, 이 상관관계를 논리적으로 해석하는 데는 무리가 있었다. 다음으로 그룹토의에서 평가한 학업관련 평가차원인 '토론능력'은 대중불안과 부적 상관관계를 보여주었고($r=-.301$, $p<.05$), '대인관계 능력(적극적 경청, 타인배려적 의사표현, 의견조율)'은 '수행불안'($r=-.375$, $p<.05$) 및 '대중불안'($r=-.302$, $p<.05$)과 부적 상관관계를 보였으며, 심리적 자원의 자기효능감($r=.311$, $p<.05$)과는 정적 상관관계를 보여주었다. 즉, 토론능력에서 높은 역량점수를 받은 참가자는 대중불안이 낮았으며, 대인관계에서 높은 역량점수를 받은 참가자들은 수행불안, 대중불안이 낮았고 자기효능감이 높았다.

한편, 프레젠테이션에서 평가한 '분석적 사고(자료이해, 통합적 사고)'와 '발표력(효과적 의사전달, 자료의 구성, 발표자세)'은 본 분석에서 사용한 변인(성격, 정서, 심리적 자원, 성적)들과 상관관계를 보여주지 않았다.

논 의

본 연구는 대입선발을 목적으로 개발된 평가센터를 이용하여, 평가센터의 신뢰도와 타당도를 확인함으로써 대학장면에서의 평가센터 활용가능성을 검증하였다.

일반화 가능도 이론의 결정연구를 이용해 신뢰도 계수(의존도 계수)를 추정한 결과, 구조화 면접이 .719, 프레젠테이션이 .854, 그룹토의가 .769로 나타나 각 평가과제가 현재 포함하고 있는 측정조건의 신뢰도는 양호한 것을 알 수 있었으며, 따라서 신뢰도 향상을 위한 추가적인 평가자나 평가차원 수의 변경은 요구되지 않았다.

평가센터의 타당도를 확인하기위해 실시한 설문조사에서 대학 재학생들은 본 평가센터에서 평가하는 역량들이 성공적인 학업수행과 적응에 중요하다고 응답하였으며, 평가센터에 참가했던 피평가자들도 본 평가센터의 공정성과 적절성에 대해 평균 4.2점 이상(5점 만점)으로 동의하는 것으로 나타났다. 이는 본 평가센터의 내용타당성과 결과타당성을 보여주는 결과라 할 수 있다.

한편, 일반화 가능도 이론의 결정연구 결과, 세 가지 평가과제 모두 수용할만한 의존도 계수를 보여주어 현재의 평가설계가 양호한 신뢰도를 보여준다고 할 수 있었다. 평가과제 국면의 수를 변화시킴에 따라 나타나는 의존도 계수의 변화추세를 살펴보면, 구조화면접의 경우 평가자의 수가 4명으로 증가하면 신뢰도 계수가 소폭 상승하지만, 2명으로 감소할 경우 보다 큰 폭으로 감소했다. 따라서 평가자의 수는 현재의 3명을 유지하는 것이 효율적으로 판단된다. 다음으로 프레젠테이션의 의존도계수 변화추세를 살펴보면 평가자 수의 증가로 인한 의존도 계수는 소폭이었으며, 평가차원의 수는 5개 이상부터는 의존도 계수에 미치는 영향이 매우 작아졌다. 결과적으로 프레젠테이션의 경우 평가자의 수가 2명, 평가차원의 수가 2개로 축소되어도 평가의 신뢰도

는 기준인 .6 이상을 유지할 수 있는 것으로 판단된다. 마지막으로 그룹토의의 의존도 계수 변화추세를 살펴보면, 다른 평가과제와 마찬가지로 평가자 수의 증가로 인한 의존도 계수는 소폭이었으나, 평가차원의 수가 증가할수록 지속적인 의존도 계수 상승을 보여주고 있다. 결과적으로 그룹토의의 경우 평가자의 수가 2명일 경우 평가차원의 수를 3개 이상으로, 평가자의 수가 3명 이상일 경우에는 평가차원의 수를 2개 이상으로 유지할 때, 신뢰로운 의존도 계수를 얻을 수 있는 것으로 판단된다.

일반화 가능도 이론의 일반화 연구의 결과, 모든 평가과제의 평균에서 평가자 국면의 영향은 주효과(r) 0.3%와 상호작용 효과(p×r) 9.9%를 합해 10.2%로 나타났으며, 평가차원의 영향은 주효과(d) 2.7%와 상호작용 효과(p×d) 18.4%를 합해 21.1%로 상대적으로 높게 나타났다. 이는 평가센터 평가가 평가자 임의대로 이루어진 것이 아니라 피평가자가 보여준 역량에 의해 이루어졌음을 의미하며, 이는 평가차원의 구성개념 타당도 증거의 하나로 볼 수 있다. 한편, 평가과제 별로 살펴보았을 때 모든 과제가 평가차원의 상호작용 효과를 보여주었으나, 평가차원의 주효과에서는 구조화면접의 경우 평가차원의 주효과가 6.1%였는데 반해, 프레젠테이션은 1.9%에 불과했고 그룹토의는 0%로 나타났다. 이는 구조화면접의 경우 질의응답이 각 역량별로 순서대로 진행되며, 한 개의 역량에 대한 질문이 끝날 때마다 해당 역량에 대한 예비평가를 하고 지원자가 퇴장 후 평가자 회의를 거쳐 역량별 최종평가를 하기 때문에 역량들이 보다 잘 변별된 것이라 할 수 있다. 그에 반해, 프레젠테이션은 발표 후에 질의응답이 역량별로 순서대로 이루어지지 않으며, 역량에 대한 평가 또한 모든 질의응답 후에 지원자가 퇴장한 후 이루어지기 때문에 역량들이 잘 변별되지 않아 평가차원 효과가 낮게 나타난 것으로 보인다. 그룹토의의 경우에도 참가자들의 30분 토론동안 평가자는 관찰만 하게 되며, 참가자 퇴장 후 역량들에 대한 평가를 하기 때문에 평가가 전반적으로 이루어져 평가차원의 주효과가 나타나지 않은 것으로 보인다.

평가차원 간의 상관관계를 분석한 결과에서는 같은 역량군에 속한 평가차원들이 상이한 평가과제에서 보여준 점수들의 상관관계(MTHM)와 다른 역량군에 속한 평가차원들이 같은 평가과제에서 보여준 점수들의 상관관계(HTMM)가 다른 역량군에 속한 평가차원들이 다른 평가과제에서 보여준 점수들의 상관관계(HTHM)보다 높게 나타나 평가과제 효과와 평가차원 효과가 모두 존재함을 보여주었다. 이러한 결과는 평가센터에 평가과제 효과와 평가차원 변별성이 모두 존재한다는 최근의 연구결과(Lievens & Conway, 2001; Bowler & Woehr, 2006, 2008)와 일관된 결과이다. 한편, 평가과제 효과를 대변하는 HTMM 상관관계(.552)가 평가차원 효과를 대변하는 MTHM 상관관계(.509) 보다 높게 나타났다는 점에서 본 평가센터가 평가차원 효과보다 평가과제 효과가 크다고 볼 수 있다. 하지만, MTHM 상관관계는 5개 상관관계가 모두 유의했었는데 반해, HTMM 상관관계는 8개 중 7개가 유의했고, 평균상관관계(.552)가 유의했던 7개 상관관계만의 평균인 점을 감안할 때, 평가과제 효과가 평가차원 효과보다 크다고 단정할 수는 없다고 할 수 있다.

평가차원과 관련변인들의 상관관계를 분석한 결과, 각 역량 별로 유의한 상관을 보여주

는 관련변인들을 살펴보면, '자기주도적 학습'과 유의한 상관관계를 보여준 변인은 학습과 관련된 내신등급 뿐이었다. '도전정신'의 경우 자기주도적 학습과는 반대로 성격에서는 외향성, 사회불안에서는 낯선 것에 대한 두려움, 심리적 자원에서는 자기효능감과 유의한 상관을 보여주었고, 교외수상실적과도 유의한 상관을 보여주었으나, 학습과 관련된 내신등급과는 상관을 보여주지 않았다. 이는 동일한 구조화면접(SI)에서 평가된 역량이지만, 자기주도적 학습과 도전정신이 서로 차별적인 역량으로 평가되었음을 말해준다고 할 수 있다. 한편, '창의성'의 경우에는 성격의 경험에 대한 개방성과 관련될 것으로 기대했으나, 오히려 외향성과 유의한 상관을 보여주었다. 이는 본 평가센터에서 창의성을 평가하는 기준이 '기존과 다른 새로운 방식'을 '주도적으로 마련'하고 '적용한 경험'으로 평가했기 때문에 창의성이 외향성과 정적 상관관계를 보여준 것으로 생각된다. 그룹토의에서 평가한 토론능력과 대인관계역량의 경우는 특히 사회불안과 상관을 보여주었다. 이는 이들 역량점수가 그룹토의 상황에서 느끼는 사회불안과 부적인 관련성이 있음을 보여주는 것이라고 생각할 수 있다. 이를 요약하면, 평가차원별로 살펴보면 학업관련 평가차원(자기주도적 학습)이 성적변인과는 관련성을 보여주었지만, 그 밖에 성격이나 정서관련 변인과는 관련성을 보여주지 않았고, 개인인성관련 평가차원들(도전, 창의, 적응)은 성적변인과는 관련성을 보여주지 않았다. 또한, 대인관련 평가차원(토론, 대인)은 특히 정서적 변인인 사회불안과의 관계에서 부적 상관관계를 보여줌으로써 평가차원들이 어느 정도 차별적으로 평가되었다고 할 수 있다. 이는 본 평가센터가 다양한 역량을 다양한 방법으로 평가함으로써, 학업수행능력 이외에도 인성 및 정서관련 잠재역량을 측정하고 있음을 시사한다.

본 연구는 대학선발 장면에 평가센터를 적용하였다는 점에서 의의가 있다고 할 수 있다. 하지만, 동시에 여러 가지 제한점을 지니고 있다. 첫째로 적은 표본수의 문제를 들 수 있다. 수행평가인 평가센터의 특성상 피평가자는 3가지 평가과제에 총 2시간 이상 참여했으며, 피평가자 1인의 평가에 평가자 3인이 배정되었다. 때문에 30인의 평가자와 3가지 평가과제 그리고 2일의 평가기간을 고려했을 때 60인 이내로 설계해야 했다. 적은 표본 수는 본 연구의 검증력(power)을 낮추었을 것으로 생각된다. 둘째는 준거관련 타당도의 문제이다. 비록 역량모델링 과정에서 준거와 관련된 역량들이 도출되긴 했지만, 대학 입학 후 학업수행이나 환경적응에 대한 실증적인 예측력은 확보되지 않았다. 본 평가센터를 통해 선발된 신입생들로부터 준거 자료를 얻기 위해서는 최소 2년 이상의 기간이 요청되기 때문이다. 더욱이 선발인원이 적기 때문에 충분한 표본수를 확보하기 위해서는 보다 오랜 기간이 요구될 것이다. 따라서 차선책으로 재학생을 대상으로 한 동시타당도 분석도 가능하다고 할 수 있으나, 이 방법 역시 평가센터 운영에 많은 비용과 시간, 그리고 많은 훈련된 평가자가 요구된다는 점에서 제약이 있다.

본 연구가 평가센터를 대입선발도구로 적용하기 위한 가능성을 확인하였으나, 전술한 바와 같이 평가센터는 비용과 시간 측면에서 부담스러운 것이 사실이다. 따라서 평가센터가 대입선발을 위한 우수한 수행평가 도구로 자리매김하기 위해, 앞으로의 연구에서는 경제성을 고려한 단축형 평가센터의 개발이 필요

하며, 그렇게 개발된 단축형 평가센터가 기존의 평가센터에 대해 지니는 기능 및 효과성에 대한 검토도 이루어져야 할 것이다. 또한, 기업조직 장면에서 발견된 우수한 준거예측력과 마찬가지로 학교장면에서도 평가센터의 준거예측력이 확보될 수 있는지에 대한 추적연구가 진행되어야 할 것이다.

참고문헌

교육과학기술부 (2007.6.14). 입학사정관제 지원계획 보도자료.

남명호 (1996). 수행평가에 있어서 일반화가능도 이론의 활용. 교육평가연구, 9(2), 73-93.

남명호 (2002). 수행평가: 기술적 측면. 서울: 교육과학사.

문혜신, 오경자 (2002). 한국판 아동·청소년용 사회 불안 척도의 타당화 연구. 한국심리학회지: 임상, 21, 429-443.

유현주 (2009). 역할에 기초한 입학사정관 전문성 훈련 프로그램. 한국교육논단, 8, 131-153.

이영식, 신상근 (2004). 다변량 일반화가능도 이론에 의한 말하기 시험의 타당도와 신뢰도에 관한 연구. *Foreign Languages Education, 11*, 249-265.

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

한국대학교육협의회 (2008). 대학입학사정관제 지원사업 실행계획.

한국대학교육협의회 (2009. 5. 4). 보도자료: 공정성 및 신뢰성 확보노력으로 입학사정관제 정착에 주력.

허창구, 신강현 (2010). 내재설계 평가센터의 신뢰도 및 타당도. 한국심리학회지: 산업 및 조직, 23, 225-249.

Bartels, L. K., Bommer, W. H., & Rubin, R. S. (2000). Student performance: assessment centers versus traditional classroom evaluation techniques. *Journal of Education for business, 75*, 198-201.

Bauer, T. N., Maertz, C. P., Dolen, M. R., & Campion, M. A. (1998). A longitudinal assessment of applicant reactions to an employment test. *Journal of Applied Psychology, 83,* 892-903.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: An examination of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478-494.

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.

Bray, D. W., Campbell, R. J. (1968). Selection of salesmen by means of an assessment center.

Journal of Applied Psychology, 52, 36-41.

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

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.

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.

Donahue, L. M., Truxillo, D. M., Cornwell, J. M., & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behavior and Personality, 12*, 85-108.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP Scales: Tiny-Yet-Effective Measures of the Big Five Factors of Personality. *Psychological Assesment. 18,* 2, 192-203.

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). *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.

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality, 12,* 13-52.

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

International Task Force on Assessment Center Guidelines (2000). Guidelines and ethical considerations for assessment center operations. *Public Personnel Management, 29,* 315-331.

Jackson, D, J., Stillman, J. A., & Englert, P. (2010). Task-Based Assessment Centers: Empirical support for a systems model. *International Journal of Selection and Assessment, 18,* 141-154.

Kottke, J. L., & Shultz, K. S. (1997). Using an assessment center as a developmental tool for graduate students: A demonstration. In R. E. Riggio & B. T. Mayes(Eds.), Perspectives on assessment centers[Special Issue]. *Journal of Social Behavior and Personality, 12*, 289-302.

Kudisch, J. D., Ladd, R. T., & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. *Journal of Social Behavior and Personality, 12,* 129-144.

Lance, C. E., Foster, M. R., Thoresen, J. D., & Gentry, W. A. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology, 89,* 22-35.

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.

Loacker, B. (1991). *Designing a national assessment system: Alverno's perspective.* National Center for Education Statistics.

Louiselle, K. G. (1980). *Confirmatory factor analysis of two assessment center rating procedures.* Paper presented at the 17th Annual IO/BO Graduate Student Conference, Minneapolis, MN.

Luthans, F., Avolio, B., Avey, J. B. & Norman, S. M. (2006). *Psychological capital: Measurement and relationship with performance and job satisfaction* (Working Paper No. 2006-I). Gallup Leadership institute, University of Nebraska _Lincoln.

Mullin, R. F., Shaffer, P. L., & Grelle, M. J. (1991). A study of the assessment center method of teaching basic management skills. In J. D. Bigelow(Ed.), *Managerial Skills: Explorations in practical knowledge,* 116-153. Newbury Park, CA: Sage.

Riggio, R. E, Aguirre, Monica, & Mayes, Bronston. (1997). The use of assessment center methods for student outcome assessment. *Journal of Social Behavior and Personality, 12*, 237-288.

Sackett, P. R., & Tuzinski, K. (2001). The role of dimensions and exercises in assessment center judgments. In M. London (Ed.), *How people evaluate others in organizations* (pp.111-129). Mahwah, NJ: Lawrence Erlbaum Associates.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87,* 735-346.

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.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149-155.

Spencer, L. M. Jr. and Spencer, S. M. (1993). *Competence at work: Models for superior performance*. New York: John Wiley & Sons.

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 : 2011. 9. 4
수정원고접수 : 2011. 10. 25
최종게재결정 : 2011. 11. 1

한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2011. Vol. 24, No. 4, 785-807

Assessment Center for Selection of University Student

Chang-Goo, Heo

The Institute of Social Science Research, Ajou University

This study were preformed for verifying the applicability of Assessment Center(AC) for selection of university student. First, the results of decision study of generalizability theory(D study) has shown acceptable reliabilities. So, we could think that this AC was designed properly. Second, the enrolled students of the university have said that the competencies rated in this AC were important for performing study and adapting to school. And participants in this AC reported they have felt fairness and they could have done their best. It means that this AC had validity. Third, generalizability study(G study) has shown dimension effect(21.1%) was higher than rator effect(10.2%). And, in the MTMM analysis, it were found both dimension effect and exercise effect. Forth, in relation analysis between AC ratings and the various records of participants, 'Self-led Study' related with 'Records of Language' positively, 'Discussion Skill' related with 'Public Anxiety' negatively, 'Challenge' related with 'Extraversion, Self Efficacy, and Record of Award off campus' positively and with 'Anxiety to unfamiliar' negatively, 'Creativity' related with 'Extraversion' positively, 'Adapting to Change' related with 'Emotional Stability' positively, and 'Interpersonal Competencies' related with 'Performance Anxiety and Public Anxiety' negatively. In short, this AC has shown applicability as selection tool for university student. Finally, the implications and limitations were discussed.

Key words : assessment center, admission officer system, generalizability, d study, g study, MTMM