한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2011. Vol. 24, No. 1, 231-255

상황판단검사의 채점용 답 결정방식과 채점방식이 타당도에 미치는 영향: 왜곡여부에 따른 비교 연구

김 의 수　　한 영 석†　　김 명 소

호서대학교 산업심리학과

본 연구에서는 상황판단검사(SJT)에서 피험자가 의도적으로 자신의 점수를 높일 수 있는지 즉, 긍정적 응답왜곡(faking good)이 가능한지를 확인하고, 만약 응답왜곡이 가능하다면 검사의 준거관련 타당도와 증분 타당도에 어떠한 영향을 미치는지 알아보고자 하였다. 이와 더불어 상이한 채점용 답 결정방식(전문가 합의, 응답자 평균, 경험적 방법)과 채점방식(Pick most 방식, 시나리오 채점방식, Best-Worst 채점방식)을 조합하여 사용하였을 경우, 응답왜곡의 정도와 검사의 준거관련 타당도 및 증분 타당도에 미치는 영향이 달라 질 수 있는지를 탐색적으로 살펴보고자 하였다. 이를 위해 국내 A대학의 리더선발 프로그램에 지원한 2학년 학생 110명과 B학과의 재학생 129명을 대상으로 적성검사, 성격검사, 상황판단 검사를 순차적으로 실시하였다. 적성검사와 성격검사는 두 집단 모두 동일한 절차와 방법으로 시행하였으나, 상황판단검사의 경우 제한된 수의 리더를 선발하기 위한 리더선발 프로그램에 지원한 학생들은 자연적으로 긍정적 응답왜곡이 발생하는 왜곡집단으로 간주하였다. B학과의 재학생들의 경우 검사의 목적이 자기진단임을 강조하고 별도의 지시문으로 솔직한 응답을 요구하였기 때문에 솔직집단으로 간주하였다. 또한 A대학의 재학생 78명을 사용하여 응답자 평균, 경험적 채점용 답을 설정하였으며, B학과의 재학생 중 성적 우수자, 리더 경험자, 대학원 1학년생을 전문가로 설정하여 전문가 기반 채점용 답으로 활용하였다. 수집된 자료들을 바탕으로 각각의 채점용 답 결정방식과 채점방식을 조합하여 솔직집단과 왜곡집단 모두에서 9개의 독립된 상황판단검사점수를 산출하였으며, 이를 바탕으로 각각의 준거관련 타당도와 증분 타당도를 비교하였다. 그 결과, 독립적으로 산출된 9개의 상황판단검사 점수 모두에서 왜곡집단이 솔직집단에 비하여 유의미하게 높은 검사점수를 보이는 것으로 나타났으며, 준거관련 타당도 또한 솔직집단이 왜곡집단보다 더 큰 것으로 나타났다. 증분 타당도의 경우 과업수행준거와 맥락수행 준거에 대한 결과가 상이하게 나타났으며, 맥락수행 준거에서만 9개의 상황판단검사 모두 솔직집단이 왜곡집단 보다 더 큰 증분적 설명량을 보였다. 이러한 결과를 토대로 본 연구의 의의와 제한점에 대하여 논의하였다.

주요어 : 상황판단검사, 응답왜곡, 채점용 답 결정, 채점방식

† 교신저자 : 한영석, 호서대학교 산업심리학과, nicehan@hoseo.edu

최근 미국 대학에서는 신입생 선발을 위하여 비인지적 평가를 도입하려는 움직임이 있다고 한다. 인지적인 부분에 초점을 두었던 기존의 선발 방식에서 벗어나 인성과 같은 비인지적 요소에 대한 평가를 포함시키겠다는 것이다(Wall Street Journal, 2009. 8. 20). 이러한 추세는 국내에서도 뚜렷하게 진행되고 있는데, 예를 들면 대학들이 최근 들어 입학사정관 제도를 도입함으로써 대입전형의 선진화를 모색하고, 인성, 리더십 등과 같은 다양한 평가요소를 제시하고 있다. 기업조직의 경우, 그 동안 채용의 중요한 잣대로 여겨지던 학력, 학점, 토익점수는 더 이상 선발장면에서 예전만큼 관심을 받지 못하고 있는 것과 맥을 같이 한다(연합뉴스, 2005. 9.26).

이처럼 다양한 장면에서 인사선발의 주요 관심사는 지원자의 인지능력에만 제한시키지 않고 비인지적 측면을 포함하는 포괄적인 평가에 주목하고 있는 실정이며, 최근 상황판단검사(SJT)는 선발장면에서 피험자들의 인지능력에 초점을 둔 전통적인 선발 검사를 대체하기 위한 중요하고 유용한 도구로 부각되고 있다(McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). 특히, 국내 기업들에서 평가센터(Assessment Center)나 상황판단검사와 같이 실제 업무 상황을 중심으로 구성된 평가도구들에 대한 관심이 증가하면서 선발 장면에서의 활용도가 높아지고 있다.

특히, 상황판단검사는 지원자들에게 직무와 관련된 가상의 시나리오를 제시한 뒤 일련의 반응대안들 중 적절한 반응을 선택하도록 하는데(Motowidlo, Dunnette, & Carter, 1990; Motowidlo, Hanson & Craft, 1997; Weekley & Jones, 1997), 활용도에 비해서 관련 연구들은 국내외적으로 부족한 실정이다. 평가 센터와는 달리 상황판단검사는 저 충실도 시뮬레이션으로서 자기보고 형식을 사용하고 있으며, 그 결과 피험자가 고의적으로 응답을 왜곡하기 쉽다는 문제점을 내포하고 있다(Lievens, Peeter, & Schollaert, 2008). 따라서, 선발 장면에서 적극적으로 활용되기 위해서는 상황판단검사에서의 반응왜곡에 대한 연구들이 선행되어야 한다. 신입사원 선발용 인성검사의 normative형식과 ipsative형식을 사용하여 응답자들의 왜곡 가능성을 비교하였던 김명소와 이헌주(2006)의 연구에서는 대학생들이 일반적으로 생각하는 주요 역량에 대한 중요도 평정점수와 현직자들이 생각하는 주요역량에 대한 중요도 평정점수는 높은 상관을 보이는 것으로 나타났다. 만약 상황판단검사가 왜곡에 영향을 받는다면, 선발 도구로서의 타당성을 위협받게 될 것이다. 이미 여러 기업들에서 선발도구로서 상황판단검사를 사용하고 있는 실정임에도 불구하고 상황판단검사에서의 왜곡에 대한 국내 연구는 전무한 실정이며, 국외의 경우도 그 수가 미미하고 연구의 결과 또한 일관적이지 못하다. 또한 이러한 연구들 중 왜곡이 검사의 준거관련 타당도 및 증분타당도에 어떠한 영향을 미치는지에 대한 연구는 매우 미흡하므로 이에 대한 종합적인 연구와 논의가 필요한 시점이다.

상황판단검사의 정의 및 특징

상황판단검사는 일련의 상황들이 지필, 언어, 또는 시각적인 형태로 제공되고 응답자의 반응이 요구되는 검사이다(Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001). 구체적으로 피험자에게 직무를 수행하는 데 있어서 경험할 수 있는 가상의 상황과 그러한 상황에

대한 대안들을 제시하고, 지원자들에게 자신이 가장 시행할 것 같은(most likely)행동이나 가장 하지 않을 것 같은 행동(least likely) 혹은 가장 효과적인 행동(best)이나 가장 최악의 행동(worst)을 선택하도록 하는 선발도구이다.

지금까지 상황판단검사를 통하여 측정하고자 하는 구성개념이 무엇이냐에 대한 많은 연구들이 있었다. 상황판단검사의 구성개념을 밝히기 위한 상위분석에서는 주로 성격 5요인(Big 5) 중 성실성(conscientiousness), 정서적 안정성(emotional stability), 원만성(agreeableness)과 인지능력을 측정하는 것으로 보고하고 있다(McDaniel, Hartman & Grubb, 2003; McDaniel & Nguyen, 2001; McDaniel et al., 2001). 하지만 Lieven, Peeters, 및 Schollaert(2007)와 McDaniel과 Nguyen(2001), Chan과 Schmitt(2005) 등은 상황판단검사에서 좀 더 다양한 구성개념을 측정할 수 있는 측정 방법이며, Weekley와 Jones(1997)는 대부분의 상황판단검사에서 직무와 관련된 기술 및 능력들의 집합체를 측정한다고 주장하였다.

이러한 상황판단검사에 대한 관심은 여러 연구결과들이 상황판단검사가 다음과 같은 장점을 가지고 있다는 것을 보여주었기 때문이다(Weekley & Polyhart, 2006). 첫째, 상황판단검사는 전통적인 선발 도구와 비교해도 만족할 만한 준거관련 타당도와 증분 타당도를 보인다. McDaniel과 그의 동료들(2001)이 수행한 상위분석에서 상황판단검사의 준거관련 타당도를 .34로 추정하였다. 이러한 수준의 타당도는 인지능력검사와 거의 비슷한 수준의 타당도 계수이다. 게다가 많은 연구들에서 상황판단검사가 인지능력검사와 성격검사와 같은 전통적인 선발도구들 이상의 증분 타당도를 갖는다는 것을 보고하였다(Clevenger, Perira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001; Weekley & Polyhart, 2005). 둘째, 상황판단검사는 인지능력검사와 비교하여 더 적은 불리효과(adverse impact)를 보인다(Motowidlo & Tippins, 1993; Pulakos & Schmitt, 1996; Weekley & Jones, 1997; Hough, Oswald, & Ployhart, 2001). 셋째, 상황판단검사의 안면타당도 때문에 피험자들은 긍정적인 반응을 보인다(Ployhart & Ryan, 1998). 이외에 저충실도 시뮬레이션으로서 상황판단검사는 평가센터와 같은 고충실도 시뮬레이션과 비교하였을 때 개발과 준비, 점수화가 쉬우며 비용과 경비가 절약되고(Motowidlo et al., 1990; Motowidlo & Tippins, 1993; Pulakos & Schmitt, 1996), 초기 선발단계에서 많은 지원자들에게 사용할 수 있다는 장점을 가지고 있다(Lieven, Peeters, & Schollaert, 2007).

상황판단검사와 가장 유사한 측정방법으로는 상황면접(Situational interview)을 꼽을 수 있다. Arthur와 Valido(2008)는 상황면접과 같은 시나리오 기반 도구들 또한 상황판단검사이며, 단지 구두형식 혹은 지필형식과 같은 시행방법에서만 차이가 있을 뿐이라고 주장한 바 있다. 상황면접은 피험자에게 직무관련 상황을 제시하고 면접관들이 피험자 반응의 효과성을 평정하는 방법이다(Latham & Saari, 1984; Latham, Saari, Pursell, & Campion, 1980). 상황면접은 형식과 타당도 측면에서 상황판단검사와 매우 유사하다(Weekley & Gier, 1987). 두 방법 사이의 차이점은 피험자에게 상황이 제시되는 형식, 피험자들이 반응하는 형식, 반응들이 점수화되는 방법에서 차이가 있다. 상황판단검사의 경우 상황이 지필형식으로 제시되는 반면에 상황면접의 경우 구두로 제시된다. 피험자들이 반응하는 형식의 경우 상황판단검사는 일련의 대안들 중에서 선택하는 방식이지만,

상황면접의 경우 구두로 반응한다. 마지막으로 상황판단검사의 경우 피험자의 반응이 몇몇의 정답 결정방법(Scoring key)과 비교되는 반면에 상황면접의 경우 면접관의 판단에 의존한다는 차이점이 있다.

상황판단검사와 같은 시뮬레이션 기법은 충실도의 차원에 따라 다양한 방법이 있을 수 있다(Ployhart, Schneider, & Schmitt, 2005). '과제자극의 충실도'(fidelity of the task stimulus)란 과제자극의 형식이 과업장면에서 마주하는 상황과 얼마나 일치하는가를 의미한다. 따라서 피험자에게 가설적인 업무 상황을 언어적으로 제시하는 지필형식의 상황판단검사는 저충실도(low-fidelity) 시뮬레이션이라고 할 수 있다. Lieven과 그의 동료들(2007)은 고충실도 시뮬레이션인 평가센터 과제와 상황판단검사의 특징을 비교하였다. 상황판단검사는 저충실도 시뮬레이션으로 지필 혹은 비디오 형식으로 자극이 제시되며 자기보고형식으로 점수는 연역적으로 결정이 되고 선별의 목적으로 대규모 지원자들에게 사용된다. 반면에 평가센터 과제는 고충실도 시뮬레이션으로서 역할연기자들의 행동적 반응 등으로 자극이 제시되며 실제 관찰 및 실제 평가 등을 통하여 점수가 산출되고 선발의 목적으로 소규모 지원자들에게 주로 사용된다.

상황판단검사에서의 응답왜곡

응답왜곡이란 호의적인 점수 및 인상을 얻기 위한 개인의 반응을 의도적으로 왜곡하는 것으로 정의될 수 있다(Dwight, 1999; McFarland & Ryan, 2000). 응답왜곡에 대한 연구는 주로 성격검사와 관련해서 수행되었는데, 연구가 연구자들 사이에서 지원자들이 응답을 왜곡하는지와 이러한 응답왜곡이 성격검사의 타당도에 미치는 영향에 대해서는 의견일치가 이루어지지 않고 있다. 몇몇 연구자들은 지원자들이 성격검사에서 응답왜곡을 하지 않으며, 설사 응답왜곡을 한다고 하더라도 검사의 타당도에 부정적인 영향을 미치지 않는다고 주장하였다(Ellingson, Smith, & Sackett, 2001; Hough, 1998; Ones & Viewsvaran, 1998). 반면에 다른 연구자들은 선발장면에서 응답왜곡이 발생하며, 성격검사의 준거관련 타당도를 감소시킨다는 것을 밝힌 바 있다(Douglas, McDaniel, & Snell, 1996; Schmitt & Ryan, 1992; Zickar, 1997). 또한 응답왜곡이 검사의 타당도에 영향을 미치지만 문제가 될 정도의 크기는 아니라고 주장하는 연구자들도 존재한다(Kamp, 1996; Paajanen, 1998).

성격검사에서의 응답왜곡 연구들과 비교하여 상황판단검사에서의 응답왜곡 연구는 그 수가 제한적이다. 또한 상황판단검사에서의 응답왜곡 연구들은 성격검사에서의 연구들과 마찬가지로 연구들 마다 상이한 결과를 보고하고 있다. 이러한 연구들은 크게 두 가지 종류로 구분할 수 있다. 첫 번째는 실험연구로서 피험자들에게 응답왜곡이 지시되었을 때, 높은 점수를 얻을 수 있는지에 초점을 두고 있다(Hooper, Cullen, & Sackett, 2006). Peeters와 Lievens(2005)의 연구에서는 솔직한 응답이 지시된 집단과 왜곡이 지시된 집단의 검사점수 사이에 유의미한 차이가 나타났으며, 효과크기(effect size)에서도 거의 1표준편차(d=.89)의 차이가 있었다. 즉, 응답왜곡이 지시된 집단과 솔직한 응답이 지시된 집단간의 상황판단검사 점수에서 유의미한 차이가 있으며, 왜곡이 지시된 집단의 검사점수가 솔직한 응답이 지시된 집단에 비하여 더 높은 점수를 보인다는

것을 밝혀냈다. 또한 이러한 왜곡이 검사의 준거관련 타당도와 증분 타당도에 부정적인 영향을 끼친다는 것을 밝혀냈다. 이와 반대로 Juraska와 Drasgow(2001)는 상황판단검사가 응답왜곡에 영향을 받지 않는다고 주장하였다. 이들의 연구에서는 상황판단검사와 성격검사에 솔직하게 응답하는 것과 기업에 고용될 수 있도록 왜곡하여 응답하는 것을 지시하였다. 그 결과, 상황판단검사에서 거의 0에 가까운 효과크기(d=.08)를 보였으며, 성격검사에서는 1표준편차가 넘는 차이가 나타났다(d=1.05).

두 번째는 현장연구로서 이러한 연구들은 실제 선발장면에서 응답왜곡이 발생되는가에 초점을 두고 있다. Reynold, Winter 그리고 Scott(1999)의 연구와 Weekley, Ployhart 및 Harold(2003), Schmidt와 Wolf(2003)의 연구에서는 현직자들의 상황판단검사 점수가 직무지원자들보다 더 높다는 것을 발견하였다(d=-.30, d=-.60, d=-.36). 또한 이들 중 Weekley, Ployhart 및 Harold(2003), Schmidt와 Wolf(2003)의 연구에서는 성격검사와 비인지적 특성을 측정하는 검사에서 직무 지원자가 현직자들 보다 더 나은 수행을 보인다고 보고하였다(d=.64, d=.75). 즉, 지원자들이 성격검사에서는 개인의 점수를 증가시킬 수 있지만, 상황판단검사에서는 그렇지 않다는 것을 의미한다(Hooper, Cullen, & Sackett, 2006). 그러나 콜센터의 현직자와 직무지원자를 대상으로 수행한 Ployhart, Weekley, Holtz 와 Kamp(2003)의 연구에서는 지원자들의 상황판단검사 점수가 현직자들의 점수보다 유의미하게 더 높은 점수(d=.88)를 보이는 것으로 나타났다.

요약하면, 상황판단검사에서의 응답왜곡에 관한 연구는 실험연구와 현장 연구 모두에서 상이한 결과들이 나타났다. 또한, 대부분의 연구들에서 상황판단검사에서의 응답왜곡이 준거관련 타당도와 증분 타당도에 영향을 주는지에 관해서는 관심을 두지 않고 있다.

상황판단검사의 채점용 답 결정방법

상황판단검사의 채점용 답 결정방법은 다양한 방식이 존재하며, 각 방법은 각기 다른 타당도 추정치를 산출할 수 있다(Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). 이뿐만 아니라 채점용 답을 결정하는 방법은 피험자 개인들의 점수를 산출하는데 핵심적인 부분으로 작용할 수 있다. Bergman과 그의 동료들(2006)은 채점용 답 결정 방식과 관련된 연구들을 조사하여 크게 경험적(empirical)방법, 이론적(theoretical) 방법, 전문가 기반(expert-based)방법으로 구분하였다. 또한 McDaniel과 Nguyen(2001)은 채점용 답 결정 방법을 주제관련 전문가(subject matter expertises: SMEs) 혹은 고성과자(excellent employee)의 합의를 통한 방법, 집단설문을 통한 방법, 경험적 방법으로 구분한 바 있다.

이러한 방법들을 하나씩 살펴보면 다음과 같다. 첫째로 경험적 방법은 문항 혹은 반응대안들이 준거와의 관계에 따라서 점수화된다(Hogan, 1994). 경험적 방법을 사용하여 채점용 답을 결정할 경우 선택지에 가중치를 부여하는 방식에 있어서 다양한 방법들이 존재한다. 이와 관련하여 박동건과 전인식(2001)은 가중치 부여 방식들 중에서 상관방식이 다른 방법에 비하여 정보의 손실이 적다고 주장하였다. 경험적 방법은 다른 채점용 답 결정 방법들과 비교하여 비교적 높은 타당도를 보인다(Hogan, 1994; Mumford & Owens, 1987). 하지만 경험적 방법은 준거의 질에 크게 의존하며 안정성 및

일반화 가능성(Mumford & Owens, 1987), 추측 가능성(Cureton, 1950)과 같은 제한점들을 가지고 있다.

두 번째는 이론적 채점용 답 결정방법으로서 상황판단검사의 정답이 특정 이론을 반영하도록 하는 방법이다. 이러한 방법에서는 상황판단검사의 문항이나 반응대안들이 이론을 반영하도록 제작되고, 이론을 반영하거나 이론과 관련된 반응대안이 정답으로 결정된다. 이론적 채점용 답 결정 방법은 경험적 방법의 제한점인 이론적 기반의 부재를 해결할 수 있으며, 일반화 가능성이 더욱 크다는 장점을 가지고 있다(Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). 하지만 피험자가 해당 이론에 대하여 잘 알고 있거나 이론 자체가 올바르지 않을 수 있다는 문제점을 가지고 있다.

세 번째로 전문가 기반 채점용 답 결정방법은 해당 주제에 대하여 많은 지식을 가지고 있는 개인들의 반응을 사용하여 채점용 답을 결정하는 방법이다. 가장 일반적으로 사용되는 전문가 기반 정답결정 방법은 주제 관련 전문가의 반응을 이용하는 방법이며 이외에 전문가의 응답과 무경험자의 응답을 비교하는 방법이 있다(Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). 첫 번째 방법은 주제관련 전문가들에게 상황과 이러한 상황에서 취할 수 있는 행동들을 제시한 뒤 각 행동의 효과성을 평정하게 하고 이에 대한 합의를 이루게 하여 정답을 결정하는 방법이다. 모든 대안을 평정하게 하는 방법 이외에도 가장 효과적인(할 것 같은) 행동과 가장 효과적이지 않은(할 것 같지 않은)행동을 선택하게 하여 정답으로 결정할 수도 있다(Motowidlo, Dunnette, & Carter, 1990). 전문가의 응답과 무경험자의 응답을 비교하는 방법은 우선 전문가들과 무경험자들 모두에게 검사를 시행하고, 전문가들이 빈번하게 선택한 반응대안들은 무경험자의 선택과 상관없이 정답으로 결정되고 무경험자들에 의하여 가장 빈번하게 선택되었지만 전문가들은 선택하지 않은 반응대안은 오답으로 결정된다. 마지막으로 상황판단검사의 채점용 답을 결정하기 위하여 집단설문 방법을 이용할 수도 있다. 이러한 방법은 몇몇의 전문가들의 응답을 사용하여 정답을 결정하는 전문가 기반 방식과 다르게 많은 사람들의 응답을 사용하여 정답을 결정하는 방법이다. 즉, 개인들에게 설문을 통하여 각 상황에 대한 반응대안의 효과성을 평정하게 한 뒤, 이의 평균값을 대안의 효과성 점수로 부여하거나 가장 효과적인(할 것 같은) / 가장 비 효과적인(할 것 같지 않은) 반응으로 가장 빈번하게 선택된 대안을 정답으로 결정하는 방법이다(Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004).

상황판단검사의 채점방식

상황판단검사의 채점용 답 결정방법 이외에 채점방식도 어떠한 방법을 사용하는가에 따라 상이한 신뢰도와 타당도를 나타낼 수 있다(Knapp, Campbell, Borman & Hanson, 2001; Waugh, 2002). Weekley와 Ployhart(2006)는 상황판단검사의 채점방식을 강제 선택형 방식과 리커트 방식으로 구분하였으며, 강민우와 윤창영 그리고 이순묵(2005)은 시나리오 채점방식과 반응대안 채점방식으로 구분하였다. 이들의 구분방법을 통합해서 각각의 방식들을 살펴보면 다음과 같다.

우선 강제선택형 / 시나리오 채점 방식 중

가장 간단한 방법은 한 개의 반응대안을 정답으로 설정하는 방법이다. 정답으로 정해진 반응대안 이외의 나머지 대안들은 모두 오답으로 처리된다. 이러한 방법에서는 문항 당 1점이 주어지며 피험자들이 선택한 반응대안이 정답일 경우 1점을 부여하며, 선택한 반응대안이 정답이 아닐 경우 점수를 부여하지 않는 방법이다.

Motowidlo와 그의 동료들(1990)은 이러한 채점방식을 확장한 채점방식을 적용하였다. 우선 전문가들과 피험자들 모두에게 문항과 반응대안을 제시하고 최고 / 가장 할 것 같은(best / most likely) 대안과 최악 / 가장 할 것 같지 않은(worst / least likely) 대안을 선택하게 한다. 그런 다음 피험자들의 응답과 전문가의 응답을 비교하여 두 개의 반응이 모두 일치하면 2점을 부여하고 한 개의 반응만이 일치할 경우 1점을 부여한다. 반대로 하나의 응답만을 반대로 선택한다면 -1점을 부여, 두 개의 응답모두 반대로 응답한다면 -2점을 부여한다. 또한 전문가들이 선택한 응답을 선택하지 않는다면 0점을 부여한다. 즉, 하나의 문항을 통하여 -2 ~ 2점의 점수를 얻을 수 있다.

또 다른 강제 선택형 / 시나리오 채점 방식은 사전에 주제관련 전문가들로부터 각 대안에 대한 평정을 받은 뒤 평균 평정값을 각 반응대안의 점수로 설정해 놓는다. 그런 다음 지원자가 선택한 최고(가장 할 것 같은)의 대안에 부여된 평균 평정값에서 지원자가 선택한 최악(가장 할 것 같지 않은)의 대안에 부여된 평균 평정값을 감하여 점수로 사용하는 방식이다. 이러한 방식은 첫 번째 방식과 달리 지원자들 간의 점수 차이를 최대로 할 수 있다는 장점이 있다(Knapp, Campbell, Borman, Pulakos, & Hanson, 2001).

한편, 리커트 / 반응대안 채점방식은 하나의 시나리오가 아니라 시나리오와 함께 제시되는 반응대안 각각을 하나의 문항으로 간주하는 방법이다. 이러한 방법에서 피험자들은 주어진 상황에 대한 각 반응대안의 효과성 혹은 자신이 할 것 같은 정도를 평정한다. 이러한 평정값들을 사전에 획득된 전문가들의 평정값과 비교하여 점수를 산출하는 방법이다. 이러한 채점방법을 더욱 구체적으로 살펴보면, 각 반응대안에 대한 전문가들의 평균 평정값에서 피험자 개인의 평정값을 감한 뒤 절대값을 점수로 활용하는 방법이 있으며(Waugh, 2002), 개별전문가들과 피험자들 사이의 평정값 차이에 대한 차이제곱 평균의 제곱근(root mean squared deviation)을 점수로 사용하는 방법도 존재한다(Wagner, 1987).

채점방식에 관한 대표적 연구인 Waugh (2002)는 미 육군 부사관의 미래 수행을 예측하고 승진에 관한 결정의 기반을 확립하기 위한 상황판단검사를 개발하였다. 응답자들은 각 응답에 대한 효과성을 평정하였으며, 각 상황에서 최선의 응답과 최악의 응답을 선택하였다. 이 연구에서는 총 6가지의 채점방식이 비교되었으며 이중 4가지 채점방식은 응답자에 의하여 선택된 최선/최악의 응답을 사용하였고, 나머지 2가지 채점방식은 각 대안에 대한 응답자의 효과성 평정점수를 사용하여 반응대안 수준에서 점수가 계산되었다. 구체적으로, 첫째 방법은 전문가 집단이 최고의 응답으로 선택한 행동을 피험자가 최고의 응답으로 선택하였을 경우 1점을 부여하는 방법이다. 두 번째 방법은 전문가 집단이 선택한 최고/최악의 응답 모두를 피험자가 최고/ 최악의 응답으로 정확히 선택하였을 경우 1점을 부여하는 방법이며, 세 번째 방법은 전문가

집단이 선택한 최고/최악의 응답과 피험자가 선택한 최고/최악의 응답을 비교하는 방법으로 점수의 범위는 -2 ~ 2점이다. 네 번째 방법은 응답자와 전문가 모두에게 각 대안의 효과성에 대한 평정을 하게 한 뒤 응답자의 평정점수에서 전문가 집단의 평균 평정점수를 감한 점수의 절대값을 점수로 사용하는 방법이다. 다섯 번째 방법은 응답자가 선택한 최고/최악의 응답에만 적용하여 점수를 계산하는 방법이고, 마지막 방법은 전문가들에게 각 대안의 효과성에 대한 평정을 하게한 뒤 응답자가 선택한 최선의 응답에 대한 전문가의 평균 평정점수에서 응답자가 선택한 최악의 응답에 대한 전문가 평균 평정점수를 감하여 점수로 사용하는 방법이다. 연구 결과, 1, 2, 3번 채점방식 간에 그리고 4, 5번 채점방식 간에 높은 상관을 보이는 것으로 나타났다. 6번 채점방식의 경우, 1 ~ 5번 방식 모두와 어느 정도의 상관이 있었고 가장 높은 준거관련 타당도를 보였다. 이러한 결과를 바탕으로 Waugh(2002)는 채점방식에 따라 검사의 신뢰도 및 타당도가 달라질 수 있다고 주장하였다.

연구문제

지금까지 살펴본 바와 같이 상황판단검사는 여러 연구들을 통하여 준거관련 타당도와 증분타당도가 입증되었다. 하지만 일반적으로 자기보고식 검사에서 문제가 될 수 있는 반응왜곡과 관련된 연구는 그 수가 제한적이며, 국내의 경우는 전무한 실정이다. 또한 대부분의 외국 연구들도 상황판단검사의 왜곡 가능성(fakability)에 초점을 두고 있으며, 이러한 왜곡이 검사의 준거관련 타당도와 증분 타당도에 어떠한 영향을 미치는지에 대하여 살펴본 연구는 Peeters와 Lievens(2005)에 수행된 연구가 전부이다. 이들의 연구에서 응답왜곡이 지시된 집단의 상황판단검사점수가 솔직하게 응답한 집단에 비하여 더 높은 점수를 보였고, 이러한 왜곡이 검사의 준거관련 타당도와 증분타당도에 부정적인 영향을 끼친다는 것을 밝혀냈다. 하지만 이러한 연구는 실제 선발장면이 아닌 실험 상황에서 의도적인 왜곡을 지시하였기 때문에 진정한 왜곡상황이라고 보기 어렵다. 따라서 실제 선발장면에서의 상황판단검사에서의 응답왜곡 여부와 이러한 응답왜곡이 준거관련 타당도와 증분타당도에 미치는 영향을 살펴볼 필요가 있다.

따라서, 본 연구는 기본적으로 Peeters와 Lievens(2005)가 수행한 연구의 반복 및 확대를 목표로 한다. 즉, 실제 선발장면에서 상황판단검사를 사용하였을 경우 지원자들이 의도적으로 반응왜곡을 할 수 있는지를 솔직한 응답이 지시된 집단에서의 상황판단검사 점수와의 비교를 통하여 알아보고, 이러한 결과가 선발결정에 영향을 미치는지 살펴보고자 한다. 또한, 만약 응답왜곡이 발생한다면 준거관련 타당도와 증분타당도에 어떠한 영향을 미치는지 파악해 볼 것이다.

그 외에도, 상황판단검사의 타당도는 채점용 답 결정 방법과 채점방식에 따라 달라질 수 있다는 기존 연구들을 기반으로 상황판단검사의 정답결정 방법과 채점 방식을 다르게 하였을 때 왜곡의 정도가 달라질 수 있는지와, 상이한 정답결정 방법과 채점방식을 사용한 상황판단검사의 준거관련 타당도와 증분타당도에 미치는 영향을 비교하고자 한다. 구체적으로, 세 가지 채점용 답 결정방식(주제관련 전문가(SME)합의 방식, 응답자 평균 방식, 경험적 방법)과 세 가지 채점방식(Best-Worst 방

식, 시나리오 방식, pick most 방식)을 조합하여 반응왜곡이 발생하는 선발장면에서 어떠한 조합이 가장 좋은 준거관련 타당도와 증분타당도를 산출하는지 탐색적으로 확인해보고자 한다.

방 법

연구대상 및 절차

국내 A대학의 리더선발 프로그램에 지원한 2학년 학생 110명(남 48명, 여 62명)과 B학과의 재학생 129명(남 42명, 여 87명)을 대상으로 적성검사, 성격검사, 상황판단 검사를 순차적으로 실시하였다. 적성검사와 성격검사는 두 집단 모두 동일한 절차와 방법으로 시행하였으나, 상황판단검사의 경우 리더선발 프로그램에 지원한 학생들을 왜곡집단으로 간주하였다. 이들에게는 상황판단검사 결과가 리더선발에 매우 중요한 비중을 차지한다고 강조함으로써 자연스럽게 왜곡을 유도하였다. 한편, B학과의 재학생들의 경우, 본 조사의 목적을 명확하게 설명하고 검사결과를 본인에게만 제시할 것을 약속하며 별도의 지시문으로 솔직한 응답을 요구하였기 때문에 솔직집단으로 간주하였다. 솔직 집단에 지시된 응답지시문은 다음과 같다.

"본 검사는 여러분 자신의 현재 모습을 진단하고 향후 개발 / 교육의 방향을 설정하기 위한 목적으로 실시되는 것이므로 솔직하게 응답하여 주시기 바랍니다."

측정도구

상황판단검사

본 연구에서 사용된 상황판단검사는 대학생용으로 Bess와 Mullins(2002)가 개발한 24문항, Ployhart와 Ehrhart(2003)가 개발한 5문항, Shuang-Yueh Pui(2007)이 개발한 6문항을 포함한 총 35문항을 국내 상황에 적합하도록 번역 및 수정하여 사용하였다. 34개의 문항은 총 4개의 반응대안과 함께 제시되었으며, 나머지 1문항은 5개의 반응대안이 제시되었다(부록 1). 응답방법은 주어진 상황에서 각각의 반응대안에 대하여 자신이 수행할 것 같은 정도를 1점 ~ 6점(1점: 가장 하지 않을 것 같은 행동 ~ 6점: 가장 할 것 같은 행동)으로 평정하게 하였다. 이와 함께 4개 혹은 5개의 반응대안 중 가장 할 것 같은 행동과 가장 하지 않을 것 같은 행동을 most와 least에 표시하게 하였다.

본 연구에서 사용된 채점용 답 결정방법은 주제관련 전문가(SME)합의 방식과 응답자 평균 방식 그리고 경험적 방법이다. 각각의 채점용 답을 산출하기 위하여 다음과 같은 방법을 사용하였다. 첫 번째 주제관련 전문가 방법은, 산업 심리학과 대학원생 4명과 학부 4학년 중 학생회장 등 학교생활에서 리더역할을 수행했던 5명에게 검사문항을 제시한 뒤, 주어진 상황에 대하여 각 반응대안이 얼마나 잘 대처하는 행동인지를 6점 척도로 평정하게 하고, 집단별로 반응대안의 효과성 점수를 합의하게 하였다. 그 다음 각 집단의 대표자들을 선정하여 집단별로 합의된 점수를 최종적으로 합의하게 하였다. 이를 바탕으로 각 높은/낮은 점수로 합의된 반응대안을 '가장 할 것 같은/하지 않을 것 같은 행동'으로 설정하였다.

두 번째 응답자 평균 방법은 A대학의 두 학과 학부생 78명(조사대상자와는 별도의 대학

생)에게 검사문항을 제시한 뒤 주어진 상황에 대하여 각 반응대안이 얼마나 잘 대처하는 행동인지를 6점 척도로 평정하도록 하였다. 그 다음 각 반응대안의 평균점수를 사용하여 '가장 할 것 같은/하지 않을 것 같은 행동'을 설정하였다.

마지막 경험적 방법은 응답자 평균방식에 참여하였던 학생들로부터 획득한 반응대안의 효과성 점수와 준거와의 상관분석을 통해 각 반응대안의 순위를 결정하였다. 이러한 순위를 통하여 '가장 할 것 같은/하지 않을 것 같은 행동'을 설정하였다.

한편, 채점 방법의 경우 Waugh(2002)의 연구에서 사용되었던 6가지 중 가장 대표적인 방법으로 구분되는 3번(Best-Worst), 6번(시나리오) 그리고 가장 간단하면서 여러 연구들에서 사용되었던 1번 방법(pick most)을 사용하여 상황판단검사 점수를 산출하였다.

성격검사

본 연구에서 사용된 성격검사는 유태용과 이기범, Ashton(2003)이 개발한 HEXACO 성격검사를 사용하였다. HEXACO 모델은 정직/겸손성(Honesty-Humility), 정서성(Emotionality), 외향성(Extraversion), 원만성(Agreeableness), 성실성(Conscientiousness), 개방성(Openness to Experience)의 총 6개 요인으로 구성되어있다. 총 문항의 수는 30문항으로 각 진술문들은 성격 특성과 관련된 행동, 사고, 기분을 나타내는 문장으로 구성되어 있으며 리커트 5점 척도를 사용하였다. 문항의 신뢰도는 요인별로 외향성은 .80, 성실성은 .62, 원만성은 .68, 개방성은 .67, 정서성은 .58 정직/겸손성은 .57로 나타났다.

인지능력 검사

피험자의 인지능력을 측정하기 위해서 모 기업에서 사용한 바 있는 선발용 적성검사의 일부분을 활용하였다. 검사는 언어력 영역과 자료해석 영역으로 구성하였으며, 문항 수는 영역 당 15문항으로 총 30문항이었다. 검사 문항은 검사 개발당시 대학생을 대상으로 시행한 예비조사에서 자료해석 영역의 경우 .76 ~ .40의 난이도와 .31 ~ .58의 변별도, 언어력 영역은 .78 ~ .40의 난이도와 .28 ~ .59의 변별도 범위를 나타냈다.

준 거

본 연구에서는 대학생들의 수행을 두 가지 측면으로 설정하였다. 첫 번째는 과업 수행으로 학업적인 측면에 초점을 두었고, 두 번째는 맥락 수행으로 조직시민행동(organizational citizenship behavior)에 초점을 두었다.

일반적으로 수행평가 장면에서 양적인 측정치로 과거 업적을 활용하고, 질적인 측면에서 동료평가, 부하평가 등 다면평가를 활용하고 있다. 이에 본 연구에서는 대학생활에서 객관적인 양적 측정치로 학점을 활용하였고, 질적인 측정치로 조직시민행동을 측정하였다. 과업 수행에 대한 측정치는 대학의 평균 학점을 사용하였으며, 자기보고식으로 조사하였다. 조직시민행동을 평가위해서는 Smith, Organ 및 Near(1983)가 개발한 조직시민행동 16문항 중 강민우와 윤창영 그리고 이순묵(2005)이 학교장면에 적합하도록 번역, 수정하였으며, '결석한 친구가 있으면 필기를 보여주고, 과제를 가르쳐 주는 등 잘 도와준다'. '수업 전 칠판을 지우거나 반장을 맡는 등 수업 진행을 위해 의무적이지 않은 일에 솔선수범한다' 등을 포함한 9문항을 사용하였다.

응답왜곡을 줄이기 위해 이들 문항에 대한 동료평가를 사용하였으며, 신뢰도는 .82로 나타났다.

분석방법

상황판단검사에서 반응왜곡이 발생하는지 알아보기 위해 왜곡집단의 검사점수와 솔직집단의 검사점수 간에 독립표본 *t*검증을 실시하여 유의미한 평균차이가 있는지 알아보았다. 또한 상황판단검사의 준거관련 타당도를 알아보기 위해서는 상관분석을 실시하였으며, 증분타당도를 알아보기 위해서는 위계적 회귀분석을 실시하였다. 모든 분석에는 SAS 9.1이 사용되었다.

결 과

왜곡집단과 솔직집단 간 차이

우선 왜곡집단에 속한 피험자들과 솔직집단에 속한 피험자들 간의 기본적인 소양 및 능력에 있어서 차이가 있는지 알아보기 위하여 평균 학점과 적성검사 점수에 대한 독립표본 *t*검증을 실시하였다(표 1). 그 결과, 평균 학점과 적성검사 모두에서 왜곡집단(*M*=3.37/10.65, *SD*=0.72/3.52)과 솔직집단(*M*=3.25/11.51, *SD*=0.54/3.79) 사이의 유의미한 차이를 보이지 않았다.

본 연구에서는 앞에서 언급한 세 가지 채점용 답 결정방식과 세 가지 채점방식을 조합하여 총 9개의 상황판단검사 점수를 산출하였다. 그런 다음 상황판단검사에서의 응답왜곡이 가능한지 알아보기 위하여 각 상황판단검사 점수들을 이용하여 독립표본 *t*검증을 실시하였다. 이와 더불어 응답왜곡의 효과가 유의미한지 알아보기 위하여 효과크기(effect size)를 계산하였다.

결과를 살펴보면, 표 2에 나타난 바와 같이, 9개 조합 모두에서 왜곡집단의 검사점수가 솔직집단의 검사점수 보다 유의미하게 더 높은 점수를 보이는 것으로 나타났다. 이를 통해서 본 검사의 선발상황에서의 왜곡이 성공적으로 조작되었다고 볼 수 있다. 또한 응답왜곡의 효과가 유의미한지 알아보기 위한 효과크기 산출 결과, 응답자 평균 채점용 답 결정방식과 pick most 채점방식의 조합(d=.72)을 제외한 모든 조합에서 1 표준편차가 넘는 효과크기를 보이는 것으로 나타났다. 그리고, 경험적 방법을 이용한 채점용 답 결정방식과 B-W채점방식을 조합하였을 때 가장 큰 효과크기(*d*=1.45)를 보이는 것으로 나타났다.

추가적으로 이러한 응답왜곡이 실제 선발결정에 영향을 주는지 알아보기 위하여 전체 응답자 중 상/하위 25%에 속하는 왜곡 응답자의

표 1. 두 집단 사이의 평균 학점 및 적성검사에 대한 t검증 결과

		왜곡집단	솔직집단	*t(df)*
평균학점	*M(SD)*	3.37 (0.72)	3.35(0.54)	0.27(199)
적성검사	M(SD)	10.65(3.52)	11.51(2.79)	-1.78(225)+

+*p*<.10

표 2. 9개 조합별 *t*검증 결과

			왜곡 집단	솔직 집단	*t*	*d*
응답자 평균	시나리오	*M(SD)*	34.85(7.93)	24.90(9.75)	8.7(236)***	1.12
	B-W	*M(SD)*	41.58(8.38)	30.21(10.63)	9.23(236)***	1.19
	Pick most	*M(SD)*	18.51(3.08)	15.98(3.87)	5.62(236)***	0.72
전문가 합의	시나리오	*M(SD)*	30.95(8.15)	21.09(9.45)	8.56(237)***	1.12
	B-W	*M(SD)*	79.54(18.41)	58.11(23.46)	7.91(236)***	1.02
	Pick most	*M(SD)*	18.31(3.55)	14.44(3.55)	8.39(237)***	1.09
경험적 방법	시나리오	*M(SD)*	25.08(9.46)	11.33(9.76)	11.02(237)***	1.43
	B-W	*M(SD)*	76.66(30.18)	32.40(31.05)	11.12(237)***	1.45
	Pick most	*M(SD)*	13.06(4.16)	8.16(3.27)	10(236)***	1.31

*** $p<.001$, ** $p<.01$, * $p<.5$, +$p<.10$

표 3. 9개 조합에 따른 상, 하위 25% 비율

		상위 25%		하위 25%	
		왜곡 응답자	솔직 응답자	왜곡 응답자	솔직 응답자
응답자 평균	시나리오	44(73%)	16(27%)	10(17%)	50(83%)
	B-W	47(78%)	13(22%)	9(15%)	51(85%)
	Pick most	41(68%)	19(32%)	10(17%)	50(83%)
전문가 합의	시나리오	48(80%)	12(20%)	10(17%)	50(83%)
	B-W	46(77%)	14(23%)	10(17%)	50(83%)
	Pick most	48(80%)	12(20%)	10(17%)	50(83%)
경험적 방법	시나리오	53(88%)	7(12%)	8(13%)	52(87%)
	B-W	52(87%)	8(13%)	7(12%)	53(88%)
	Pick most	53(88%)	7(12%)	10(17%)	50(83%)

비율과 솔직 응답자의 비율을 알아보았다. 분석 결과, 채점용 답 결정방식과 채점방식의 조합에 따른 9개의 상황판단검사 모두에서 상위 25%에서는 왜곡 응답자가 솔직 응답자보다 더 많이 포함되어 있는 것으로 나타났으며, 하위 25%에서는 왜곡 응답자 보다 솔직 응답자가 더 많이 포함되어 있었다(표 3).

위의 결과를 종합해보면 상황판단검사에서

는 피험자들이 의도적으로 자신의 검사점수를 높이기 위해 긍정적으로 왜곡하여 응답할 수 있으며, 이러한 응답왜곡은 선발결정에 있어 영향을 미친다는 것을 알 수 있다.

응답왜곡이 준거관련 타당도에 미치는 영향

응답왜곡이 검사의 준거관련 타당도에 미치는 영향을 알아보기 위하여 세 가지 채점용 답 결정방식과 세 가지 채점방식에 따라 총 9가지의 상황판단검사와 준거 측정치와의 상관분석을 실시하였고, 이는 표 4에 제시되어 있다.

먼저, 과업수행 준거인 평균 학점과 응답자 평균 방식을 사용하여 채점용 답을 설정한 상황판단검사의 준거관련 타당도를 살펴보면, 솔직집단에서 시나리오, B-W, Pick most 채점방식 각각 .31, .29, .25로 유의미한 상관을 보이는 것으로 나타났으며, 왜곡집단의 경우도 각각 .22, .22, .21로 유의미한 상관을 보이는 것으로 나타났다. 상관의 크기는 솔직집단이 모든 방식에서 왜곡집단의 상관보다 더 큰 것으로 나타났지만, 상관차이 검증 결과 유의미한 차이를 보이지는 않았다.

전문가 합의 방식을 사용하여 채점용 답을 설정한 상황판단검사의 준거관련 타당도를 살펴보면, 솔직집단의 경우 시나리오, B-W, Pick most 채점방식 각각 .26, .32, .18로 모두 유의미한 상관을 보이는 것으로 나타났지만, 왜곡집단의 경우 .18, .19, .08로 B-W방식에서만 유의미한 것으로 나타났다(Fisher's z =-2.2, p< .05). 응답자 평균 방식과 마찬가지로 솔직집단의 상관이 왜곡집단의 상관보다 모두 큰 것으로 나타났지만, 상관차이 검증 결과 유의미한 차이를 보이지 않았다.

또한, 경험적 방식을 사용하여 채점용 답을 설정한 상황판단검사의 준거관련 타당도를 살펴보면, 솔직집단의 경우 시나리오, B-W, Pick most 채점방식 각각 .26, .27, .25로 모두 p=.05

표 4. 9개 조합에 따른 준거관련 타당도

		왜곡집단			솔직집단		
		M(SD)	GPA	OCB	*M(SD)*	GPA	OCB
응답자 평균	시나리오	34.85(7.93)	.22*	.22+	24.9(9.75)	.31***	.34***
	B-W	41.58(8.38)	.22*	.20	30.21(10.63)	.29***	.37***
	Pick most	18.51(3.08)	.21*	.10	15.98(3.87)	.25**	.31**
전문가 합의	시나리오	30.95(8.15)	.18	.22+	21.09(9.45)	.26**	.40***
	B-W	79.54(18.41)	.19*	.25*	58.11(23.46)	.32***	.41***
	Pick most	18.31(3.55)	.13	.09	14.44(3.55)	.18*	.33***
경험적 방법	시나리오	25.08(9.46)	.08	.25*	11.33(9.76)	.26**	.45***
	B-W	76.66(30.18)	.13	.26*	32.4(31.05)	.27**	.43***
	Pick most	13.06(4.16)	-.02	.15	8.16(3.27)	.25**	.38***

GPA(학점): 과업수행준거, OCB(조직시민행동): 맥락수행준거

*** p< .001, ** p < .01, * p < .05

수준에서 유의미한 상관을 보이는 것으로 나타났지만, 왜곡집단의 경우 .08, .13, -.02로 모두 유의미한 상관을 보이지 않았다. 위의 결과들과 마찬가지로 솔직집단의 상관크기가 왜곡집단의 상관크기보다 모두 큰 것으로 나타났으며, 상관차이 검증 결과 Pick most 채점방식에서만 유의미한 차이가 있었다($z=-2.09$, $p<.05$).

다음으로, 맥락수행 준거인 조직시민행동 측정치와 응답자 평균방식을 사용하여 채점용답을 설정한 상황판단검사의 준거관련 타당도를 살펴보면, 솔직집단의 경우 시나리오, B-W, Pick most 채점방식 각각 .34, .37, .31로 유의미한 상관을 보이는 것으로 나타났으나, 왜곡집단의 경우 각각 .22, .20, .10으로 시나리오 채점방식에서만 $p<.10$ 수준에서 유의미한 상관을 보였다. 상관의 크기는 솔직집단이 모든 방식에서 왜곡집단의 상관보다 더 큰 것으로 나타났으며, 상관차이 검증 결과 모두 유의미한 차이를 보이지는 않았다.

전문가 합의 방식의 경우에는 솔직집단이 시나리오, B-W, Pick most채점방식에서 각각 .40, .41, .33으로 모두 유의한 상관분석 결과를 보였다. 한편왜곡집단의 경우 .22, .25, .09의 상관이 도출되었는데, 시나리오 채점방식은 $p<.10$의 수준에서, B-W채점방식에서는 $p<.05$ 수준에서 유의미한 상관을 보였다. 상관의 크기에 있어서도 솔직집단이 왜곡집단 보다 맥락수행 측정치와 더 큰 상관을 보이는 것으로 나타났으며, 상관차이 검증에서는 모두 유의미한 차이를 보이지 않았다.

마지막으로 경험적 방법을 사용하여 채점용답을 설정한 상황판단검사에서는 솔직집단이 시나리오, B-W, Pick most채점방식 각각에서 .45, .43, .38의 상관이 도출되어 모두 $p<.001$의 수준에서 유의미한 것으로 나타났다. 반면에 왜곡집단의 경우에는 .25, .26, .15으로 시나리오 채점방식과 B-W 채점방식에서만 $p<.05$수준에서 유의미한 상관을 보였다. 상관의 크기에 있어서도 솔직집단이 왜곡집단 보다 모든 채점방식에서 더 큰 상관을 보였으나, 상관차이검증 결과 모두 유의미한 차이를 보이지 않는 것으로 나타났다.

위의 결과를 종합해 보면, 과업수행 준거와 맥락수행 준거 모두에서 솔직집단의 준거관련 타당도가 왜곡집단의 준거관련 타당도 보다 높은 것으로 나타났다. 이는 상황판단검사에서의 응답왜곡이 검사의 준거관련 타당도에 부정적인 영향을 미친다는 Peeters & Lievens (2005)의 연구를 지지하는 결과이다. 상관차이 검증에서는 과업수행 준거에서 경험적 채점용답 결정방법과 Pick most 채점방식을 조합하였을 때만 유의미한 상관차이가 있었다. 또한, 과업 수행에서는 전문가 합의 방식과 B-W 채점방식을 조합하였을 때가 가장 높은 준거관련 타당도를 보였으며, 맥락수행에서는 경험적 방식과 시나리오 채점방식을 조합하였을 때 가장 높은 준거관련 타당도를 보이는 것으로 나타났다.

응답왜곡이 증분 타당도에 미치는 영향

다음으로 준거관련 타당도에서와 마찬가지로 응답왜곡이 상황판단검사의 증분 타당도에 어떠한 영향을 미치는지 알아보기 위해 위계적 회귀분석을 실시하였다. 우선 과업수행을 예측하기 위한 위계적 회귀분석에서 적성검사, 성격검사, 상황판단검사가 순차적으로 투입되었다. 분석 결과, 솔직집단에서는 적성검사와 성격변인들이 각각 준거변인의 11%, 8%를 유

의미하게 설명하는 것으로 나타났다. 상황판단검사의 추가적 설명량은 채점용 답 결정방식과 채점방식의 조합에 따라 1% ~ 4%의 범위를 보이며 모두 유의한 영향을 미치는 것으로 나타났다. 한편, 왜곡상황의 경우 적성검사와 성격변인들은 각각 9%와 12%의 설명량을 보이는 것으로 나타났으며 상황판단검사의 추가적인 설명량은 채점용 답 결정방식과 채점방식의 조합에 따라 0% ~ 5%의 범위로 영향을 미치는 것으로 나타났다.

다음으로 맥락수행을 예측하기 위한 위계적 회귀분석에서도 과업수행을 예측하기 위한 위계적 회귀분석과 동일한 절차를 사용하여 실시하였다. 결과를 살펴보면 솔직집단에서는 적성검사와 성격변인들이 각각 준거변인의 4%, 18%를 유의미하게 설명하는 것으로 나타

표 5. 과업수행에 대한 적성검사, 성격검사, 9개 조합 상황판단검사의 위계적 회귀분석

		솔직집단			왜곡집단		
		B	R^2	$\triangle R^2$	B	R^2	$\triangle R^2$
STEP1							
적성검사		.33***	.11***		.30**	.09***	
STEP2							
			.19***	.08***		.21***	.12***
외향성		-.06			-.16		
성실성		.15			.16		
원만성		-.09			-.04		
개방성		.15			.08		
정서성		.08			.20*		
정직겸손성		.22*			.14		
STEP3							
응답자 평균	시나리오	.19	.22***	.03***	.24*	.26***	.05***
	B-W	.16	.21***	.02**	.23*	.25***	.04***
	Pick most	.14	.21***	.02**	.21*	.25***	.04***
전문가 합의	시나리오	.17	.21***	.02***	.18	.23***	.02***
	B-W	.22*	.23***	.04***	.15	.23***	.02***
	Pick most	.12	.20***	.01**	.11	.22***	.01**
경험적 방법	시나리오	.20*	.22***	.03***	.09	.22***	.01**
	B-W	.16	.21***	.02***	.13	.22***	.01***
	Pick most	.18*	.22***	.03***	-.02	.21***	-

*** $p < .001$, ** $p < .01$, * $p < .05$

났다. 상황판단검사의 추가적 설명량은 채점용 답 결정방식과 채점방식의 조합에 따라 7% ~ 16%의 범위를 보이며 모두 유의한 영향을 미치는 것으로 나타났다. 왜곡상황의 경우 적성검사와 성격변인들은 각각 0.1%와 11%의 설명량을 보였으며, 상황판단검사의 추가적인 설명량은 채점용 답 결정방식과 채점방식의 조합에 따라 0% ~ 2%의 범위를 보이는 것으로 나타났다.

위의 결과들을 종합해 보면 과업수행 준거에서는 전문가 합의/B-W 조건, 경험적방법/시나리오, 경험적방법/B-W, 경험적방법/Pick Most 조건에서는 솔직집단이 더 높은 증분적 설명량을 보이는 것으로 나타났다. 응답자 평균 답결정 방식에서는 모든 조건과, 전문가합의/시나리오, 전문가합의/Pick Most조건에서는 왜

표 6. 맥락수행에 대한 적성검사, 성격검사, 9개 조합의 위계적 회귀분석

		솔직집단			왜곡집단		
		B	R^2	$\triangle R^2$	*B*	R^2	$\triangle R^2$
STEP1							
적성검사		.21+	.04+		-.03	.001	
STEP2							
			.22**	.18**		.11	.11
외향성		.24*			.34*		
성실성		.20+			.03		
원만성		.05			.16		
개방성		-.11			-.04		
정서성		-.23*			.01		
정직겸손성		-.21			.05		
STEP3							
응답자 평균	시나리오	.19	.25**	.07**	.11	.12	.01
	B-W	.32**	.30***	.12***	.13	.12	.01
	Pick most	.37***	.33***	.15***	.15	.13	.02
전문가 합의	시나리오	.23+	.26**	.08**	.09	.12	.01
	B-W	.33**	.30***	.12***	.18	.13	.02
	Pick most	.38***	.34***	.16***	.16	.13	.02
경험적 방법	시나리오	.20+	.25**	.07**	-.02	.11	-
	B-W	.21+	.26**	.08**	-.03	.11	-
	Pick most	.30**	.30***	.12***	.04	.11	-

*** $p < .001$, *** $p < .01$, ** $p < .05$, + $p < .10$

곡집단의 증분설명량이 높은 것으로 나타났다. 한편, 맥락수행 준거에서는 9개 조건 모두에서 솔직집단이 더 높은 증분적 설명량을 보이는 것으로 나타났다. 과업수행 준거에 대한 증분 타당도 분석결과는 일관적인 결과를 나타내지 않았지만, 맥락수행 준거에서는 솔직집단과 왜곡집단의 증분적 설명량은 상당한 차이를 보였다.

논 의

최근 다양한 장면에서 상황판단검사에 대한 관심과 활용도가 증가되고 있으나 상황판단검사의 응답왜곡 가능성에 관한 연구들은 미흡한 실정이며, 특히 국내의 경우 전무하다. 또한 대부분 선행연구들은 응답왜곡이 가능한지에 초점을 두고 있었으며, 검사의 타당도에 미치는 영향에 대해서는 거의 다루지 않고 있었다. 따라서 본 연구의 목적은 상황판단검사에서의 응답왜곡이 가능한지 알아보고 이러한 응답왜곡이 선발결정과 검사의 타당도에 어떠한 영향을 미치는지 알아보는 것이다. 또한 상이한 채점용 답 결정방법과 채점방식을 조합하여 사용하였을 때 일관적인 결과가 나타나는지 탐색적으로 살펴보았다.

연구 결과, 9개의 검사점수 모두에서 응답집단과 솔직집단 사이에 유의미한 차이가 존재하였으며, 기대한 바와 같이 왜곡집단이 솔직집단 보다 더 높은 점수를 보였다. 또한 상황판단검사점수의 상위 25%와 하위 25%를 살펴보았을 때, 9개 조합 모두에서 상위 25%에는 왜곡 응답자가, 하위 25%에는 솔직 응답자가 더 많이 포함되어 있는 것으로 나타났다. 이를 통하여 상황판단검사에서 응답왜곡이 가능하며, 이러한 응답왜곡은 선발결정에 심각한 영향을 줄 수 있다고 해석할 수 있다.

검사의 준거관련 타당도에 미치는 영향을 알아보기 위한 상관분석결과, 과업수행 준거와 맥락수행 준거 모두에서 솔직집단의 준거관련 타당도가 더 큰 것으로 나타났으며, 이는 9개 조합 모두에서 일관적인 결과를 산출하였다. 이를 통하여 상황판단검사에서의 응답왜곡이 검사의 준거관련 타당도에 부정적인 영향을 미치는 것으로 결론내릴 수 있으며, 이는 Peeters와 Lievens(2005)의 연구를 지지하는 결과이다.

또한, 응답왜곡이 검사의 증분 타당도에 미치는 영향을 알아보기 위한 위계적 회귀분석을 실시했을 때 과업수행 준거와 맥락수행 준거에서 일관적인 결과를 보이지 않았다. 우선 과업수행 준거에서는 9개의 상황판단검사 중 전문가합의/B-W 조합, 경험적 방법/시나리오, 경험적 방법/B-W, 경험적 방법/Pick most 조합을 사용하였을 경우에만 솔직집단이 더 높은 증분적 설명량을 보이는 것으로 나타났다. 반면 맥락수행의 경우 9개 조합 모두에서 솔직집단이 왜곡집단 보다 더 높은 증분적 설명량을 보였다.

이와 같은 상이한 결과는 상황판단검사에서 사용된 지시문의 영향일 수 있다. 상황판단검사의 지시문은 지식형(Knowledge) 지시문과 행동경향(Behavior Tendency)지시문으로 구별할 수 있다. 지식형 지시문의 경우 피험자들에게 주어진 상황에서 가장 효율적인 혹은 비효율적인 행동을 선택하게 하는 것이며, 행동경향 지시문은 주어진 상황에서 가장 할 것 같은 행동과 하지 않을 것 같은 행동을 선택하게 하는 형식이다. 기존 연구들에서 지식형 지시문은 지원자의 인지적인 측면과 더욱 관련

이 있으며, 행동경향 지시문은 성격변인들과 더욱 관련이 있다고 알려져 있다(McDaniel, Hartman, Whetzel, & Grub, 2003; McDaniel. & Nguyen, 2001). 본 연구에서는 행동경향 지시문을 사용하였기 때문에, 두 가지 종류의 준거 중 맥락수행 준거는 지원자의 행동경향이 중요한 예측변인이 될 수 있다. 또한 준거관련 타당도의 결과를 살펴보면, 과업수행 준거와 상황판단검사의 관련성 보다 맥락수행과 더 큰 관련성을 보였는데, 이러한 결과도 지시문의 영향을 뒷받침 해줄 수 있다. 따라서, 맥락수행 준거에 대한 결과만을 보면, 상황판단검사에서의 응답왜곡은 검사의 증분 타당도에 부정적인 영향을 미친다고 결론 내릴 수 있다.

본 연구는 상황판단검사에서의 응답왜곡이 검사의 준거관련 타당도와 증분 타당도에 미치는 영향을 살펴본 Peeters와 Lievens(2005)의 연구의 제한점을 보완하고 확장하여, 임의로 지시된 응답왜곡 상황이 아닌 실제 선발장면에서의 상황판단검사의 응답왜곡 가능성을 알아보았다는 점에서 의의가 있다. 이와 더불어, 연구자 임의로 채점용 답 결정방식과 채점방식을 결정하였던 기존의 연구를 확장하여 다양한 채점용 답 결정방식과 채점방식을 조합하여 적용하였을 경우에도 일관적인 결과가 나타나는지를 살펴본 점에서도 학문적 의의가 있다고 할 수 있다.

또한 본 연구에서는 상황판단검사의 준거관련 타당도가 지지되었는데, 과업수행 준거와는 .32(솔직집단), 맥락수행 준거와는 .45(솔직집단)로 유의미한 상관계수가 산출되었다. 이와 같은 수준의 준거관련 타당도는 McDaniel과 그의 동료들(2001)이 수행하였던 상위분석에서의 결과와 유사하거나 더 높은 수준의 타당도 계수이다. 이를 통하여 직무수행을 예측하기 위한 선발도구로서 상황판단검사의 효용성을 입증한 점에서 기업의 인사담당자들에게 유용한 정보를 제공한 것으로 평가할 수 있다. 증분 타당도의 경우 솔직집단의 과업수행준거와 맥락수행 준거에서 상이한 패턴의 결과를 도출하였지만, 상황판단검사의 지시문의 효과를 감안하였을 때, 기존의 선발도구들 이외에 추가적으로 최대 18%의 설명량(솔직집단)을 보이는 것으로 나타났다. 이러한 결과는 응답왜곡이 최소화되는 상황에서 상황판단검사가 사용된다면, 타당한 선발도구로 추가될 수 있음을 시사해준다.

또한 본 연구는 상황판단검사가 자연적으로 응답왜곡이 발생하는 실제 선발장면에서 사용될 경우, 검사의 준거관련 타당도와 증분 타당도에 부정적인 영향을 미칠 수 있다는 사실을 밝혀냈다. 비록 상황판단검사의 준거관련 타당도의 경우 왜곡집단에서도 어느 정도 수용할 만한 수준의 타당도 계수를 보였지만, 증분 타당도의 경우 솔직집단과 왜곡집단 사이의 차이가 매우 큰 것으로 나타났다. 이미 많은 기업들이 적성검사 및 성격검사와 함께 상황판단검사를 사용하고 있는 상황에서 이러한 결과는 새로운 선발도구로서의 상황판단검사의 효용성을 위협할 수 있기 때문에 인사담당자들이 유념해야 할 정보이다. 따라서 기존의 선발도구들과 함께 사용하기 위해서는 피험자들의 응답왜곡을 방지하기 위한 방법이 반드시 강구되어야 할 것이다. 본 연구의 결과를 고려하였을 때, 피험자들에게 솔직한 응답을 요구하는 별도의 지시 즉, 왜곡 응답에 대한 경고가 한 가지 방법이 될 수 있다.

또한, 자연적으로 응답왜곡을 통제할 수 없는 상황이라면 준거관련 타당도 측면에서 전

문가합의/B-W 조합이 가장 타당한 조합이라고 판단된다. 왜냐하면, 본 연구 결과 이 조합에서 왜곡집단과 솔직집단 모두에서 평균학점 및 조직시민행동과 모두 유의한 상관값을 보였고, 과업수행에 대한 증분타당도 분석결과에서 솔직집단의 경우 가장 높은 증분 설명량을 나타냈으며 왜곡집단에서도 유의한 증분 설명량을 보였기 때문이다. 이는 Knapp, Campbell, Borman, 및 Pulakos(2001)가 수행한 연구에서 채점방법에 따른 상황판단검사의 신뢰도를 비교하였는데, 전문가들의 평균 평정값과 피험자들이 응답한 가장 효율적/비효율적인 답의 차이값을 활용한 채점방법에서 가장 높은 신뢰도를 보고한 연구결과와 맥을 같이 한다.

선발 장면에서 상황판단검사를 이미 사용하고 있거나 혹은 사용할 예정인 기업의 인사선발 담당자들에게 검사 실시 시 응답 왜곡에 대한 경고와 함께 정답 결정방식은 전문가 합의 방식을 사용하고, 채점방식은 B-W 방식을 사용하여 결과를 도출하는 것이 바람직하다는 시사점을 제공한 점에서 본 연구의 실용적 의의를 찾을 수 있다. 물론 B-W 방식에서 검사 구성 시에 대안들의 바람직성 조사를 통해 유사한 수준의 바람직성 대안을 배치하는 것의 고려해야할 것이다.

본 연구는 몇 가지 측면에서 제한점을 가지고 있다. 첫째, 본 연구에서는 대학생들을 대상으로 실시하였기 때문에 실제 기업의 선발 장면으로 연구결과를 일반화시키기에는 한계점을 가지고 있다. 따라서 추후의 연구에서는 실제로 기업에서 사용되는 상황판단검사에서의 응답왜곡과 응답왜곡이 검사의 타당도에 어떠한 영향을 미치는지에 대하여 알아볼 필요가 있다. 둘째, 본 연구의 결과는 한 대학의 재학생들을 대상으로 실시하였기 때문에 더욱 범위의 제한과 관련된 문제가 있을 수 있다. 따라서 앞으로 대학 및 기업조직 장면에서 다양한 표본을 사용하여 일관된 결과가 나타나는지를 확인해보아야 할 것이다. 셋째, 본 연구에서는 행동경향 지시문을 사용하였지만 또 다른 지시문의 형태인 지식형 지시문을 사용하는 상황판단검사에서는 응답왜곡이 선발결정과 검사의 타당도에 어떠한 영향을 미치는지와 상이한 채점용 답 결정방법과 채점방식을 조합하여 적용하였을 때 어떠한 패턴의 결과가 나타나는지 알아볼 필요가 있다. 마지막으로 본 연구에서는 대학생의 수행을 두 가지 측면 즉, 과업수행과 맥락수행으로 구분하였지만 이외에 다양한 측면에서의 수행이 존재할 수 있다. 따라서 추후의 연구는 다양한 측면에 대한 준거를 확보하여 종합적인 관점에서 수행되는 것이 바람직하겠다.

참고문헌

강민우, 윤창영, 이순묵 (2005). 지시문과 채점방식에 따른 상황판단검사의 타당도 비교. 한국심리학회지: 산업 및 조직, 18(3), 547-565.

김명소, 이헌주 (2006). 성격검사의 형식이 응답왜곡에 미치는 효과. 한국심리학회지: 산업 및 조직, 19(3), 371-393.

박동건, 전인식 (2001). 전기자료(biodata) 문항의 가중치 부여 체계간의 타당도 연구: 분석집단 크기에 따른 비교연구. 한국심리학회지: 산업 및 조직, 14(1), 101-113.

연합뉴스 (2005). 신입사원 채용 시 인, 적성검사 강화. 9월 26일자.

이순묵 (2003). 지필형 상황판단검사에 대한 비평적 고찰: 한국 심리학회지: 산업 및 조직, 16(3), 129-154.

유태용, 이기범, Ashton, M. C. (2004). 한국판 HEXACO 성격검사의 구성타당화 연구. 한국심리학회지: 사회 및 성격, 18(3), 61-75.

Arthur, W., & Valido, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93,* 435-442.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. (2006). Scoring situational judgement tests: Once you get the data, Your trouble begin. *International Journal of Selection and assessment*, *14,* 223-235.

Bess, T. L., & Mullins, M. E (2002). Exploring a dimensionality of situational judgment: task and contextual knowledge. Paper presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Canada.

Chan, d., & Schmitt, N (2005). Situatioanl judgment Tests. In A. Evers, N. Anderson, & O. Smit-Voskuijl(Eds.) The blackwell handbook of Personnel selection(pp.219-242). The blackwell Publisher.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmidt-Harvey, V. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, *86,* 410-417.

Clevenger, J. p, & Haaland D. E. (2000). Examining the relationship beetween job knowledge and situational judgment performance. Paper presented at the 15th Annual Conference of the Society of Industrial and Organizational Psychology. New Orleans. April.

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational Measurement*, *5,* 115-124.

Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996). The validity of non-cognitive measure decays when applicants fake. Paper presented at the annual conference of the Academy of Management, Cincinnati, OH.

Dwight, S. A. (1999). An assessment of the difference of warning applicants not to fake. Unpublished doctoral dissertation, State University of New York at Albany.

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the Influence of social desirability on personal factor structure. *Journal of Applied Psychology*, *86,* 122-133.

Hogan, J. B. (1994). Empirical keying of background data measures. In G.S. Stokes, M.D. Mumford and W. A. Owens(Eds), Biodata handbook: Theory, research, and use of biographical information in selection and performance predictions(pp.69-107). Palo Aloto: Consulting Psychology Press.

Hooper, A. C., Cullen, M. J., & Sackett, P. R. (2006). "Operational threat to the use of SJTs: faking, coaching, and retesting issue", in Weekley, J. A., & Ployhart, R. E.(Eds), Situational Judgment Tests: Theory, Measurement and Application, Lawrence Erlbaum Associates, Mahwah, NJ, pp.205-232.

Hough, L. M. (1998). Effects of intentional

distortion in personality measurement and evaluation of suggested palliatives. *Human Performance*, *11*, 209-244.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Detrerminats, detection, and amelioration of adverse impact in personnel selection procedure: Issue, evidence, and lessons learned. *Internal Journal of Selection and assessment*, *9,* 152-194.

Juraska, S. E., & Drasgow, F. (2001). Faking situational judgment: A tests of Conflicts Resolution Skills Assessment. Paper presented at the 16th annual conference of the Society for Industrial and Organizational Psychology, San Diego. CA.

Knapp, D. J., Campbell, C. H., Borman, W. C.,Pulakos, E. D., & Hanson, M. A. (2001). Performance assessment for a population of jobs. In J. P. Campbell & D. J. Knapp(Eds), Exploring the limits in personnel selection and classfication. Mahwah, N. J: Lawrence Erlbaum Associates.

Latham., G. P., & Saari, L. M. (1984). Do people do what they say? Further studies of the situational interview. *Journal of Applied Psychology*, *69*, 569-573.

Latham., G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of atplied psychology*, *65*, 422-427.

Lievens, F., Peeters, H & Schollaert, F. (2008). Situational judgment tests: a review of recent research. *Personnel Review. 37.* 426-411.

McDaniel, L. A, Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgement test to predict job performance: A clarification of the literature. *Journal of Applied psychology*, *86*, 812-821.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & GrubⅢ, W. L. (2003). Situational Judgments, response instructions, and validity: A meta analysis. *Personnel psychology*, *60*, 63-91.

McDaniel., M. A & Nguyen, N. T. (2001). Situational judgment tests: a review of practice and construct assessed. International *Journal of Selection and assessement*, *9*, 103-113.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across non-cognitive measure: Effects on faking behavior and test measurement properties. *Journal of Personality Assessment*, *78l*. 348-369.

Motowidlo, S. J. & Tippins, N. (1993). Further studies of low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology*, *66*, 337-344.

Motowidlo. S. J., & Dunnet, M.D & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied psychology*, *75,* 640-647.

Motowidlo. S. J., Hanson. M. A., & Craft. J.L (1997). Applied measurement method in industrial Psychology. In D. I., Whetzel & G. R. Wheaton(Eds). Applied measurement method in industrial psychology(pp.241-260). Palo Alto. CA: Davies-Black

Mumford, T. V., & Owens, W. A. (1987). Methodology review: Principles, procedures,

and findings in the application of background data measures. *Applied Psychological Measurement*, *11*, 1-31.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, *11*, 245-269.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. j., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*, 187-208.

Paullin, C., & Hanson, M. A. (2001). Comparing the validity of rationally-derived and empirically-derived scoring keys for situational judgment inventory. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college student's success: the influence of faking. *Educational and psychological Measurement*. *65*, 70-89.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11,* 1-16.

Ployhart, R. E., & Ryan, A. M. (1998). The relative importance of procedure and distributive justice in determining applicant's reactions. *Journal of applied psychology*, *83,* 3-16.

Ployhart, R. E., Schnider, & Schmitt, N. (2005). Organizational staffing: Contemporary practice and theory. Mahwah, NJ: Lawrence Erlbaum Associates.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology*, *56*, 733-752.

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance*, *9*, 241-258.

Reynolds, D. H., Winter, J. L., & Scott, D. R. (1999). Development, validation and translation of a Professional-level situational judgment Inventory items. Invited presentation to College Board, New York.

Schmidt, D. B., & Wolf, P. P. (2003). Susceptibility of SJTs to applicant faking: An examination of applicant and incumbent samples, Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Shuang-Yueh Pui. (2007). Situational Judgement Tests: A Measurement of Judgement?. The degree of Master of Arts in Industrial and Organizational Psychology the Graduate College of Bowling Green State University.

Smith, K. C., & McDaniel, M. A. (1998). Criterion and construct validity evidence for a situational judgment measure. Paper presented

at the 13th annual conference of the Society for Industrial and Organizational Psychology, Inc., Dallas, TX.

Smith, C. A.,. Organ, D W & Near, J (1983). Organizational Citizenship Behavior: Its Nature and Antecedents. *Journal of Applied Psychology, 68*, 653-663.

Wall Street Journal. (2009). Adding Personality to the College Admissions Mix. 8월 20일자

Waugh, G. (2002). Selecting response options and item for situational judgment test. Paper presented as part of following symposium - Understanding and predicting Performance in future jobs. 17th Annual conference of *society for industrial and organizational psychology*, Tronto.

Weekley, J. A., & Jones, C. (1997). Wideo-based situational testing. *Personnel psychology*, *50*, 24-49.

Weekley, J. A., & Ployhart, R. E. (2005). Situational judgment: Antecedents and relation ship with performance. *Human Performance*, *18*, 81-104.

Weekley, J. A., & Ployhart, R. E. (2006). An introduction to situational judgment testing. In J. A, Weekley & R. E. Ployhart(Eds.) *Situational judgment tests*. Mahwah, NJ: Lawrence Erlbaum Associates.

Weekley, J. A., Ployharts, R. E., & Harold, C. (2003). Personality and situational judgment tests across applicant and incumbent setting: An examination of validity, measurement, and subgrouping differences. Paper presented at the 18th Annual Conference of the *Society for Industrial and Organizational Psychology*, Orlando, FL.

Weekley., J. a. & Gier, J. A. (1987). Reliability and validity of situational interview for a sales position. *Journal of applied Psychology*, *72*, 484-487.

Zickar, M. J. (1997). Computer simulation of faking on a personality test. Paper Presented at the annual conference of the *Society for Industrial and Organizational Psychology*, St Louis, Mo.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis, *Journal of Applied Psychology*, *84,* 551-563.

1차 원고접수 : 2011. 1. 10
2차 원고접수 : 2011. 2. 14
최종게재결정 : 2011. 2. 18

한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2011. Vol. 24, No. 1, 231-255

Comparison of validities for scoring keys and scoring algorithms in situational judgment test: the influence of faking

Eui Soo Kim Young Seok Han Myoung So Kim

Hoseo University

The purpose of the present study was to examine the fakability of the situational judgment test. Specifically, the study was focused on the following questions; (1) whether participants are able to fake their answers on the situational judgment test in the real situation of selection, (2) whether faking influences the criterion-related validity of the situational judgment test and its incremental validity over cognitive and personality tests, and (3) whether the combination of different scoring key(SME consensus, average in response, and empirical keying) and different scoring algorithm(scenario, Best-Worst, and Pick most) has influence on the degree of fakability as well as both criterion-related validity and incremental validity of the situational judgment test. 110 students who applied to the leadership program were considered the faking group, while 129 students of B department at A university were considered the honest group. The members of both groups completed a cognitive test, a personality questionnaire and a situational judgment test. Only for the situational judgment tests, each group was asked to respond as instructed. Another group of 78 students of A university participated in the survey to develop two scoring key(empirical, average in response keying). SME consensus key was developed by 9 SMEs(5 undergraduate students with leadership and good GPA, 4 graduate students). And then 9 situational judgment scores were produced independently. Results indicated that the all scores of students in the faking group were significantly higher than those of students in the honest group. Furthermore, criterion-related validity of the situational judgement test in the honest group was higher than that of the faking group for both task performance and contextual performance. While faking had negative effects on the criterion-related validity for both criteria of performance, incremental validity of the situational judgement test in the honest group was higher than that of the faking group only for the contextual criteria. Finally, the limitation and future direction of the present study were discussed.

Key words : Situational judgment tests, faking, scoring key, scoring algorithm

부록 1. 상황판단검사 예제 문항

14. 이번 학기 수강하는 과목들 중 매우 어려운 과목의 시험 날짜가 정해졌는데, 마침 그 날까지 제출해야 하는 과제가 있다면 당신은 어떻게 하시겠습니까?

		동의도	Most	Least
①	과제를 내주신 교수님과 상담하고 과제 마감일을 연장해주실 것으로 요청 드린다.	① ② ③ ④ ⑤ ⑥		
②	불이익을 감수하고 시험 끝난 후 바로 과제를 시작하여 늦게 제출한다.	① ② ③ ④ ⑤ ⑥		
③	제출일 전날 밤까지 과제를 완성한다. 과제의 질을 포기하더라도 마감시간에 맞춰 제출한다.	① ② ③ ④ ⑤ ⑥		
④	마감기한 며칠 전에 미리 과제를 완성하여 시험 전날에는 충분히 시험 공부를 할 수 있도록 한다.	① ② ③ ④ ⑤ ⑥		

15. 당신은 이번 학기 수강하는 수업들 중 한 수업에 어려움을 느끼고 있습니다.
중간고사를 망쳤기 때문에 얼마 남지 않은 기말고사에서 좋은 성적을 받아야 한다는 부담감이 큽니다.
당신은 대학에 들어온 후 열심히 공부하였고 지금까지 우수한 학점을 받아 왔는데,
이 수업 때문에 학점이 낮아지는 것을 원하지 않습니다. 어떻게 하겠습니까?

		동의도	Most	Least
①	계속해서 공부를 하고 설사 이해하기 어려운 개념들이 나와도 포기하지 않는다.	① ② ③ ④ ⑤ ⑥		
②	교수님을 찾아가 잘 모르는 부분을 질문하고 이 과목을 공부하는 법에 대해서 조언을 받는다.	① ② ③ ④ ⑤ ⑥		
③	중간고사 때 공부했던 것보다 더 많은 시간을 이 과목의 공부에 할애한다.	① ② ③ ④ ⑤ ⑥		
④	모든 과목에서 좋은 학점을 받을 수 없다는 것을 인정하고, 당신이 현재 잘 하고 있는 다른 과목들에 시간과 노력을 집중한다.	① ② ③ ④ ⑤ ⑥		