한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2009. Vol. 22, No. 1, 53-70.

지식형 상황판단검사의 채점용 답(Scoring key) 결정과 채점 방식에 따른 준거관련 타당도 비교 연구

정 현 선 박 동 건†

고려대학교 심리학과

본 연구는 상황판단검사의 채점용 답 결정 방식(SME 합의, 응답자 평균, 경험적 결정 방식)과 채점 방식(-2~2, B-W 방식)의 조합에 따라 산출된 점수의 유사성과 준거 관련 타당도가 어떠한 양상으로 나타나는지를 지식형 리더십 상황판단검사를 통해 탐색적으로 접근하여 비교하고자 하였다. 국내 A기업에서 근무하는 과장 이상의 재직자들을 대상으로 상황판단검사의 문항 개발, 채점용 답 결정 단계를 거쳐, 과장급 이상의 395명에게 검사가 실시되었다. 그 결과, 채점용 답 결정과 채점 방식 조합에 따라 산출된 점수들 중 SME × B-W의 조합이 가장 높은 준거 관련 타당도를 보이는 것으로 나타났다. 본 연구 결과로 볼 때, 상황판단검사의 채점용 답을 어떤 방식으로 결정하는가가 검사의 측정적, 예측력 측면에서 중요한 이슈임을 파악할 수 있다.

주요어 : 상황판단검사, 채점용 답(Scoring key) 결정, 채점방식, 리더십

---

† 교신저자 : 박동건, 고려대학교 심리학과, sykhpark@korea.ac.kr
익명의 심사위원 세 분의 상세한 지적에 감사드립니다.

많은 연구자들과 기업들은 조직의 선발 절차를 개선하기 위해 노력해왔으며, 이러한 과정에서 직무 지원자들이 하는 판단 혹은 결정의 질을 평가하고자 하는 다양한 시도들이 있어왔다(Chan & Schmitt, 2005). 이는 실제 직무를 수행하는 과정에서 판단과 반응을 취해야 하는 다양한 상황에 부딪히게 되고, 판단과 반응의 적절성 여부는 결과적으로 직무 수행의 성공 여부로 직접적으로 연결된다는 점에서 그 중요성이 더욱 강조되어 가고 있기 때문이다. 이러한 맥락에서 일반적으로 조직에서 업무상 중요하게 작용하였던 사건들에 대해 응답자가 어떠한 판단을 할 것인가를 바탕으로 개인의 역량을 평가하고자 하는 상황검사의 일종인 상황판단검사에 대한 관심 또한 그와 함께 최근 더욱 증가하고 있다(예, Chan & Schmitt, 2002; Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel & Nguyen, 2001; Motowidlo, Dunnette, & Carter, 1990; Ployhart & Ehrhart, 2003). 그 동안 상황판단검사의 타당도나 불리효과(adverse impact) 측면에 대한 긍정적인 주장들은 여러 경험적 연구들을 통해 확인되어 왔다. McDaniel, Morgeson, Finnegan, Campion, 그리고 Braverman (2001)이 실시한 상위분석(meta analysis)에 따르면, 상황판단검사의 준거관련 타당도(ρ)는 .34로 나타났으며, 응답자들의 보다 솔직한 응답을 유도하는 효과를 볼 수 있음이 확인되었다. 또한 인지 능력검사(cognitive ability tests)와 비교할 때 불리효과(adverse impact)가 상대적으로 적게 나타나며(Pulakos & Schmitt, 1996; Motowidlo & Tippins, 1993), 외견상의 안면타당도를 넘어, 직무관련 영역을 나타내는 내용타당도도 지니고 있는 것으로 나타났다(Chan & Schmitt, 1997; Motowidlo, Hanson, & Crafts, 1997; Salgado, Viswesvaran, & Ones, 2001). 이에 따라 현재 미국과 유럽의 많은 기업에서 상황판단검사를 유용한 선발도구로 사용하고 있다(McDaniel, Hartman, Whetzel & Grubb Ⅲ, 2007; McDaniel 등, 2001; Salgado 등, 2001).

Motowidlo, Hanson 그리고 Crafts(1997)가 상황판단검사의 전형적인 개발 단계를 기술한 이후, 상황판단검사에 대한 관심과 필요성에 대한 인식의 증가에 부합하여 상황판단검사 개발에 대한 다양한 시도들이 있어왔다. 하지만 다양한 시도들이 있어온 만큼 검사 개발과정에서 적용되는 측정 방법들도 여러 연구자들에 의해 상이한 패턴으로 적용되어 왔으며, 결과적으로 다양한 유형의 상황판단검사가 개발되어 사용되어 오고 있다. McDaniel 등(2001)은 102개의 상황판단검사 관련 연구들을 바탕으로 한 상위분석 연구에서 검사 점수와 준거성과점수와의 관계에 영향을 미치는 조절변인이 존재할 것이며, 이는 검사의 유형과 개발 특성에 연관되어 발생할 수 있다는 주장을 하였다. 이들 연구자들의 주장은 우리가 상황판단검사를 보다 잘 이해하기 위해서는 검사의 유형과 개발 과정에 대한 보다 깊은 이해가 요구된다는 것을 암시하고 있다. 실제 조절효과를 분석한 결과, 직무 분석을 바탕으로 개발된 검사들이 그렇지 않은 검사들보다 더 큰 타당도 계수를 갖는 것으로 나타났으며(.38 vs .29)(McDaniel 등, 2001), 또 다른 연구에서는 상황판단검사의 지시문의 형태에 따라 구성개념 타당도와 준거관련 타당도가 달라지는 것으로 나타났다(McDaniel 등, 2003). 이처럼 지금까지 상황판단검사의 유형과 특성에 따라 측정적인 특성이 다르게 나타나는지 여부를 확인하기 위한 시도들이 있어 온 것은 사실이

다. 하지만 아직까지 소수 연구들에 그치고 있으며 상황판단검사의 점수를 산출하는데 가장 핵심적인 부분이 되는 채점용 답(scoring key)을 결정하는 방식이 검사의 측정적 특성에 어떠한 영향을 미칠 수 있을 것인가를 파악할 수 있는 연구는 거의 이루어 지지 않고 있다. 일반적으로 채점용 답 결정 방식은 주제관련 전문가 집단을 통해 산출하는 방식과 응답자 표본 평균을 통한 방식, 전기자료 채점용 답 산출 방식을 활용하는 방법 등 일반적으로 이러한 세 가지의 채점용 답 결정방식이 제안되어 왔다(McDaniel & Nguyen, 2001). 상황판단검사를 개발하고 있는 연구자들은 제안 된 방식들 중 하나를 선택하여 채점용 답을 산출하고 있지만, 이들 채점용 답 결정 방식들을 직접적으로 비교한 경험적 연구는 거의 이루어지지 않고 있다. 상이한 방식을 사용하더라도 산출한 상황판단검사의 결과가 동일한가 여부가 파악되지 않은 상태에서 연구자마다 채점용 답 결정방식을 각각 상이하게 사용할 경우, 이를 통해 얻을 수 있는 정보는 제한적일 수 밖에 없을 것이다.

이러한 맥락에서 본 연구는 상황판단검사의 점수를 산출하는데 가장 핵심적인 요소인 채점 정답, 즉 채점용 답을 산출하는 방식을 상이하게 적용하였을 때, 검사의 준거 관련 타당도가 상이하게 나타나는지를 리더십 상황판단 검사를 사용하여 탐색적으로 살펴 볼 것이다. 또한 널리 사용되는 두 가지의 채점 방식과 함께 연결하여 이들 조합에 따른 준거 관련 타당도의 차이 여부를 비교하고, 가장 높은 준거 관련 타당도를 나타내는 조합은 무엇인지를 확인해 볼 것이다.

상황판단검사(Situational Judgement Test)의 전형적 개발 과정

상황판단검사의 개발 및 사용에 관한 것은 1920년대까지 거슬러 올라 갈 수 있으나, 인사 선발에서 예언 변인으로서의 상황 판단에 대한 초기 토대가 되었던 검사들로는 1940년대 이후에 개발된 실용판단검사(practical judgment test)(Cardall, 1942) 감독수행검사(Supervisory Practices Test) (Bruce & Learner, 1958), 및 감독판단검사(Supervisory Judgment Test)(Greenberg, 1963)를 들 수 있다. 검사명에서도 알 수 있듯이, 이들 검사들은 응답자가 다양한 행동들의 바람직성을 평가하거나 혹은 특정 상황에 대응하는 행동 대안들 중에서 선택하는 것을 필요로 하는 관리(supervisory) 판단력에 관련된 것들이다. 가장 현대적인 상황판단 검사의 형태는 Motowidlo, Dunnette 그리고 Carter(1990)의 작업에서 비롯되었다(Chan & Shmitt, 2005). Motowidlo 등(1990)은 직무분석을 사용하여 특정 직무에서의 판단 유형을 규명하고, 검사의 내용, 안면 타당도를 향상시키는 방식으로 상황판단검사를 개선하였으며 이는 관리자, 의사, 군인 등 특정 직무 혹은 집단을 표적으로 하는 현재의 다양한 상황판단검사로 이어져 오고 있다. 일반적으로 상황판단검사의 개발 과정은 Motowidlo, Hanson 그리고 Crafts(1997)가 제안한 방식을 전형적으로 따르고 있는데, 그 절차는 다음의 단계들로 설명이 될 수 있다. 우선 상황판단 검사가 표적으로 하는 작업 관련 구성개념을 파악하고, 이 구성개념과 연관되는 능력 혹은 전문성을 필요로 하는 상황 혹은 중요 사건을 직무종사자가 직접 작성하거나 설문이나 면접을 통해 검사 개발자가 구성을 하게 된다. 그 다음, 다른 직무 종사자들에게 이러한 상황을 해결하거나

개선하는데 요구되거나 나타날 수 있는 다양한 행동 반응들을 기술하도록 하여 가능한 행동 대안들을 수집한다. 수집된 반응 대안들은 또 다른 직무 종사자들에게 해당 상황에서 가장 좋은 것(best)과 제일 나쁜 것(worst)을 선택하도록 하거나 혹은 그들의 상대적인 효과성을 평정하도록 하게 되는데, 이러한 과정을 통해 검사에 최종적으로 포함될 행동 반응들을 선별하고, 채점용 답(scoring key)이 결정된다. 마지막으로 구성된 검사 문항들을 바탕으로 응답자들에게 하나의 시나리오에 대해 행동 대안들 중, 가장 할 것 같은(most likely) 행동과 가장 하지 않을 것 같은(least likely) 행동(혹은 가장 좋은, 가장 좋지 않은 대안)을 선택하도록 하여 검사를 실시하게 된다. 최종 상황판단검사 점수는 응답자들이 선택한 반응에 이전 단계에서 얻어진 채점용 답을 적용하여 산출된다(Chan & Schmitt, 2005).

대부분의 상황판단검사 연구자들은 기본적으로 이러한 절차에 준하여 검사를 개발해 오고 있지만, 각 국면의 세부 과정에서 연구자들 간에 적용 방식의 차이가 나타나고 있다. 그러한 차이들 중, 상황판단검사의 측정적인 유형이 좌우될 수 있는 핵심적인 국면으로는 채점용 답을 결정하는 과정, 채점용 답을 적용하여 채점을 하는 과정 그리고 응답자들에게 요구하는 응답 반응이 무엇인가, 즉 검사의 지시문에 대한 결정 과정을 들 수 있다. 그것은 상황판단검사의 이들 세 과정 국면에서 연구자가 어떠한 방식을 선택하였는가에 따라 개인이 받게 되는 검사 점수가 상이하게 나타날 수 있으며, 결과적으로는 검사의 심리측정적 특성 역시 좌우될 수 있는 부분들이기 때문이다.

상황판단검사의 지시문 유형

실제로 상황판단검사 연구자들은 검사 개발 과정에서 응답자들의 반응을 요구하는 다양한 방식들을 시도하고 있다. Motowidlo 등(1997)이 제시한 방법에서는 응답자들이 자신이 하게 될 반응과 가장 가까운 것과 가장 먼 것을 각각 고르는 방식을 사용하였다. 반면, Chan과 Schmitt(2002), Clevenger, Pereira, Wiechmann, Schmitt 그리고 Schmidt(2002)의 연구들에 기술된 상황 판단 검사에서는 가능한 대안 각각을 효과성 측면에서 5점 척도로 평가하도록 하는 방식을 사용하였다. 또한 McDaniel과 Nguyen (2001)은 가장 좋은 안(best)과 가장 좋지 않은 안(worst)을 선택하도록 하는 방식을 사용하였다. 이처럼 지시문의 형태는 검사에 따라 다르게 제시될 수 있는데, McDaniel 등(2003)은 지시문의 형태를 크게 행동경향(behavioral tendency)과 지식형(knowledge) 지시문으로 구분하고, 이에 따른 검사 특성의 구분을 시도하였다. 행동경향형은 해당 상황에서 자신이 “가장 할 것 같은 행동/하지 않을 것 같은 행동(most/least likely)”을 묻는 것으로, 응답자에게 어떻게 행동할 것인지를 묻고 있기 때문에 의도와 행동의 일관성 즉, 행동경향성(behavior tendency)과 관련이 클 것이라고(McDaniel & Nguyen, 2001; Motowidlo et al., 1990) 설명을 하고 있다. 반면 지식형은 해당 상황에서 취할 수 있는 최선/최악(best/worst)을 선택하도록 하는 것으로, 어떤 것이 최선인지 최악인지를 구분하는 적절성에 대한 판단 및 지식 즉, 인지적 능력을 요구하는 형태라고(McDaniel & Nguyen, 2001) 설명하고 있다. 이들의 연구에서 두 가지 방식을 비교한 결과, 지식형의 지시문이 보다 높은 준거관련 타당도를 보이

는 것으로 나타났다. 또한 McDaniel과 Nguyen (2001)은 지식형의 지시문이 응답자들에게 가장 좋은 대안을 선택하거나 각 응답에 효율성을 평정하도록 지시하기 때문에 행동경향성 질문에 비해 사회적 바람직성에 의한 반응 왜곡의 가능성이 줄어들게 될 것이라 주장하였다.

상황판단검사 채점용 답 결정 방식

일반적으로 상황판단검사의 채점용 답을 결정하는 방식은 크게 3가지 방식으로 구분을 할 수 있다(McDaniel & Nguyen, 2001). 우선은 주제관련 전문가들(subject matter expertises: SMEs) 혹은 고성과자(excellent employee)에게 상황에서 취할 수 있는 여러 반응대안들의 효과성을 평정하도록 한 후, 이에 대한 합의(consensus) 과정을 통해 결정된 점수를 각 행동반응들의 채점용 답으로 설정하는 방법이 있다. 또한 이는 주제관련 전문가들에게 각 행동 반응들의 효과성을 모두 평정하도록 하는 것이 아니라 가장 효과적인 안과 가장 효과적이지 않은 안을 각각 하나씩 선택하도록 하여 가장 좋은 안과 그렇지 못한 안으로 채점용 답을 확정하는 방식에도 적용이 될 수 있다(Motowidlo 등, 1990).

두 번째로는 소수의 고성과자나 주제관련 전문가들의 논의를 통한 합의 방식이 아니라, 집단 설문을 통해 정답을 구성할 수 있다. 즉, 설문을 통해 개인들의 반응 대안에 대한 효과성을 평정 받은 후, 중심 경향값 즉, 평균을 각 대안의 효과성 점수로 부여를 하거나 혹은 가장 효과적/효과적이지 않은 반응으로 많이 선택된 반응들을 각각 최적/최악의 반응으로 채점용 답을 확정하는 방법이 사용될 수 있다(Oswald, Schmitt, Kim, Ramsay & Gillespie, 2004). Oswald 등(2004)은 연구에서 집단 설문을 통한 채점용 정답 개발에서 고학년 대학생들을 활용하였으며, 이 역시 일종의 학업 장면에서 주제관련 전문가 집단을 활용한 방식이라 할 수 있을 것이다. 하지만 준거에 근거한 고성과자 중심으로 구성된 것이 아니라는 점에서 볼 때 경험이 있는 응답자 표본을 활용한 것에 가깝다 볼 수 있다.

마지막은 경험적 방식으로, Mumford와 Whetzel(1997)이 제안한 전기자료(biodata)에 대한 채점용 답 결정방식들을 활용하는 것이다(McDaniel & Nguyen, 2001). Mumford와 Whetzel (1997)는 여러 전기자료 채점 점수 부여 방식들 중 상관, 회귀 가중(regression weighting), 그리고 WAB(weighted application blank) 방식을 제시하고 있는데, 본 연구에서는 전기자료의 선택지별 가중치 부여 방식들 중 상대적으로 정보 손실이 적은(박동건 & 전인식, 2001) 상관 방식을 사용하고 있다. 기존의 연구에서 주제관련 전문가 집단의 합의를 통해 채점용 답을 산출하는 방식과 응답자 집단의 평균 활용하는 방식에 따른 검사 점수의 비교를 시도한 상황판단검사 연구는 아직 없는 상황이며, 반면 응답자 표본과 전기자료 방식을 차용한 경험적 방법을 비교한 Parker, Golden, Russell 그리고 Redmond(2000)의 연구에서는 응답자 표본 활용 방식이 보다 적절한 방법인 것으로 나타났다.

상황판단검사의 채점 방식

상황판단검사의 점수를 산출하기 위한 채점 방식은 매우 다양하게 나타나고 있으며, Waugh(2002)는 여섯 가지 가능한 대표적인 채

점 방식을 제시를 하였다. 첫 번째는 응답자에게 최선의 대안을 선택하게 하고, 선택한 반응이 주제 전문가에 의해 정해진 최선의 대안과 일치할 경우에 1점을 부여하는 방식이다. 두 번째는 최선과 최악의 대안을 모두 제대로 선택하면 1점을 부여하며, 세 번째는 Weekly & Jones(1999)에서 사용된 것과 같이 사전에 설정된 채점용 답에 준하여 최선과 최악의 선택 각각 올바르게 하였을 경우에 각각 1점씩을 부과하고, 정해진 최선/최악을 반응을 최악/최선의 반응으로 선택하였을 경우에는 각각 -1점씩을 부여하여 -2에서 2점을 부여하는 방식이다. 네 번째는 사전에 설정된 행동반응의 효과성 점수와 지원자의 반응평정의 차이 절대값을 사용하는 것이며, 다섯 번째는 최선과 최악의 대안에 대해서만 반응 평정의 차이 절대값 계산하는 방식이다. 마지막으로 여섯 번째 방식은 응답자가 선택한 최선과 최악의 반응에 행동 반응 대안들 각각에 부여된 효과성 점수 적용하여, 응답자가 선택한 최선의 반응에 부여되어 있는 채점용 답 점수에서 최악의 반응에 부여된 채점용 답 점수를 빼는 것이다. 이 방식은 사전에 정해진 효과성 점수가 가장 높은 반응과 가장 낮은 반응을 각각 선택할수록 높은 점수를 받게 된다. Waugh(2002)는 이 여섯 가지 방식을 통해 산출한 점수들을 분석한 결과, 채점 방식에 따른 점수는 서로 동등(equivalent)하지 않으며, 결과적으로 채점 방식이 신뢰도와 타당도에 영향을 주게 된다고 주장을 하였다. 또한 Knapp, Campbell, Borman, Pulakos 그리고 Hanson(2001)은 이들 여섯 가지 방식들 중에서 크게 구분되는 가장 대표적인 방식은 세 번째와 여섯 번째 방식으로 여섯 번째 방식이 다른 방식에 비해 가장 많은 정보를 제공해 주며, 동시에 가장 높은 준거 관련 타당도를 나타내고 있다고 보고를 하였다.

연구문제

이처럼 상황판단검사는 그 개발과정에서 지시문과 채점용 답 결정 방식, 그리고 채점방식을 무엇으로 선택하였는가에 따라 다양한 유형으로 나타날 수 있으며, 이를 통해 얻어진 검사 점수 역시 동일하지 않게 나타날 수 있을 것이다. 본 연구에서는 검사의 지시문이 지식형 혹은 행동 경향형인가에 따라 측정하고 있는 구성개념이 상이할 수 있다는 McDaniel 등(2003)의 연구를 바탕으로 리더십이라는 구성개념을 측정하기 위해 보다 적절한 지시문의 유형이 무엇인가를 확인하기 위한 사전 연구를 실시하였다. 사전에 개발된 58개 문항을 사용하여 과장급 이상의 재직자 120명을 대상으로 실시한 결과, 행동경향형과 비교할 때 지식형, 즉 상황에 대응하기 위한 반응 행동 대안들 중 가장 최선의 대안과 최악의 대안을 고르도록 하는 방식이 리더십을 측정하는데 상대적으로 좋은 신뢰도(.60 vs .52)와 준거 타당도(.21 vs .17)를 보이는 것으로 나타났다. 따라서 본 연구에서는 지식형 지시문으로 구성된 상황판단검사 점수가 채점용 답 결정 방식과 채점 방식의 조합에 따라 상이하게 나타나는지, 각 조합으로 산출된 점수간의 관계성과 준거관련 타당도에 있어 차이가 있는지 여부를 비교하고, 이 때 어떠한 조합으로 개발이 되었을 때 가장 좋은 준거관련 타당도를 보이는지를 탐색적으로 확인하는 것을 연구문제로 설정을 하고 있다. 본 연구는 세 가지 채점용 답 결정 방식과 두 가지를 채점용 답 결정 방식을 조합하여, 6개의

상황판단검사를 산출, 그에 따른 검사 점수 특성과 준거 관련 타당도를 비교할 것이며, 그 중 준거를 보다 잘 예측하기 위한 최적의 조합은 무엇인가에 대해 살펴볼 것이다. 본 연구에서 적용, 비교할 세 가지 채점용 답 결정 방식은 주제관련 전문가들의 합의를 활용하는 방법, 응답자 표본 평균의 평균을 활용하는 방법 그리고 준거와 관계성을 활용하여 경험적인 결정을 하는 방식이다. 반면 채점 방식은 Knapp 등(2001)이 가장 대표적인 방식으로 제안한 두 가지 방식을 적용하고 비교할 것이다. 한 가지 방식은 사전에 설정된 채점용 답에 준하여 최선과 최악 반응을 올바르게 선택하였을 경우에 각각 1점씩을 부과하고, 정해진 최선/최악을 반응을 최악/최선의 반응으로 선택하였을 경우에는 각각 -1점씩을 부여하여 -2에서 2점을 부여하는 방식(-2~2 방식)이다. 또 다른 하나는 행동 반응 대안들 각각에 부여된 효과성 점수 적용하여, 응답자가 선택한 최선의 반응에 부여되어 있는 채점용 답 점수에서 최악의 반응에 부여된 채점용 답 점수를 빼는 방식(B-W 방식)을 사용할 것이다.

방 법

표본 및 연구 절차

본 연구에서는 중공업, IT, 주류, 전자, 출판 등의 계열사를 보유하고 있는 국내 A 대기업에서 근무하는 과장 이상의 재직자들을 대상으로 연구를 실시하였다. 연구는 상황판단검사 문항을 개발하는 단계와 상황판단검사 채점용 답을 결정하는 단계, 그리고 마지막으로 완성된 검사를 과장 이상의 395명을 대상으로 검사를 실시하여 채점용 답 결정 및 채점 방식을 각각 적용하여 산출한 개인들의 점수를 분석하는 단계로 진행이 되었다.

측정도구

상황판단검사

본 연구에서 사용한 리더십 상황판단검사는 전략적 비전 제시, 효율적 업무 조율, 결단 및 실행, 그리고 변혁적 성과 관리 네 개의 차원으로 구성이 되었다. 전략적 비전 제시는 리더로서 조직원들에게 건설적인 조직의 나아가야할 방향을 제시하고 이를 실행할 수 있는가를, 효율적 업무 조율 차원은 조직원의 역할 관리 및 책임 배분, 보상 및 지권, 인력자원 활용에 대한 측면을 다루고 있다. 또한 결단 및 실행은 리더로서 결단력, 추진력 등에 대한 개인의 역량을, 마지막으로 변혁적 성과 관리는 조직원들의 수행에 대한 모니터링(monitoring), 피드백(feedback), 적절하고 공정한 기회 제공의 리더 역량에 관련된 차원이다. 이러한 4개의 차원에 준해 검사 문항이 개발되었으며, 문항의 개발 절차는 Motowidlo 등(1990)이 제안한 절차에 따랐다. 우선 문항에 사용된 상황을 구성하기 위해 중요사건기법(critical incident technique)(Flanagan, 1954)을 이용하여 중요사건을 수집하였으며, 이는 기업 재직자들을 대상으로 실시한 설문과 면접을 통해 수집되었다. 이 과정을 통해 총 78개의 상황이 개발되었으며, 이에 대한 행동 반응 대안을 개발하기 위해 과장급 이상의 재직자 100명에게 개방형 설문을 실시하여 가능한 행동 반응 대안들을 수집하였다. 수집된 행동 반응 대안들은 빈도, 내용 측면을 고려하여 선별 한 후, 좋은 대안부터 좋지 않은 대안을

고르게 포함시키기 위해 과장급의 고성과자 50명을 주제관련 전문가로 별도로 선정하여, 각 행동 대안에 대한 효과성 점수를 논의하여 평정하도록 하였다. 이를 바탕으로 가장 좋은 행동 반응에서 가장 좋지 않은 반응까지의 대표적인 5개의 반응 선택지를 최종 구성하였으며 검사 개발과정에 참여하지 않은 과장급 이상 120명을 대상으로 58개 문항을 사용하여 예비 검사를 실시하였다. 예비 검사에서는 행동경향형과 지식형 지시문 검사가 모두 사용되었으며, 최종적으로 지식형을 지시문으로 한 34문항의 검사지가 구성이 되었다.

채점용 답 결정 방식

주제관련 전문가(SME) 합의 방식. 주제관련 전문가 합의 방식의 채점용 답을 결정하기 위해 과장급 이상의 고성과자 40명을 주제관련 전문가로 선정하여 워크숍을 실시하였다. 고성과자는 회사에서 제공한 성과자료를 바탕으로 해당 기업의 인사 담당자가 설정을 하였다. 채점용 답은 4~5명으로 구성된 10개의 조를 크게 5개의 조씩 두 개 집단으로 다시 구분하여, 각각 검사의 50%에 해당하는 문항들에 대한 반응대안의 효과성 정도를 10점 척도로 평정하도록 하였다. 최종 행동 반응의 효과성 점수는 조별로 합의된 점수를 다시 전체 5개의 조가 최종적으로 합의하여 점수를 결정하는 방식을 사용하였다. 이를 바탕으로 가장 높은/낮은 점수로 합의된 반응을 최적/최악의 대안으로 설정을 하였다.

응답자 평균 방식. 120명을 대상으로 한 예비검사에서 지식형 검사를 부여받은 65명의 응답자들에게 각 문항에 제시된 5개의 행동 대안들을 조직 효과성 측면에서 10점 척도로 평정하도록 지시를 하였다. 최종 채점용 답은 그 평균점수로 사용하였다. 그 다음 평균점수를 바탕으로 가장 높은/낮은 평균점수 나타난 반응을 최적/최악의 대안으로 설정하였다.

준거 관련성 기반 경험적 결정 방식. 예비 검사에서 응답자들에게 받는 반응들의 효과성 점수와 준거의 상관분석을 통해 각 행동 반응들의 순위를 결정을 하였으며, 그 순위에 따라 관계성이 정적인 값으로 가장 크게 나온 것을 최적의 대안으로, 부적으로 가장 큰 절대값을 갖는 것은 최악의 대안으로 설정하였다. 이 때 사용한 준거는 응답자에 대한 관리능력에 대한 수행 평가 점수를 사용하였으며, 5개의 각 반응에 대한 효과성 점수는 순위에 따라 가장 높은 것을 10점으로 하고 가장 낮은 순위를 보인 것을 2점으로 하여, 10~2점까지의 점수를 순차적으로 부여를 하였다.

채점 방식

-2~2 방식. 앞서 설명한 바와 같이 사전에 설정된 채점용 답에 준하여 최선과 최악의 선택 각각 올바르게 하였을 경우에 각각 1점씩을 부과하고, 정해진 최선/최악의 반응을 최악/최선의 반응으로 선택하였을 경우에는 각각 -1점씩을 부여하여 -2에서 2점까지 가능하도록 하였다. 따라서 응답자가 세 가지의 채점용 답 결정 방식에 따라 설정된 최선/최악의 반응 대안들과 일치하게 최선/최악으로 선택을 하였을 경우 2점을 부여하였고, 최선/최악의 반응 대안을 각각 최악/최선의 반응 대안으로 선택을 하였을 경우에는 -2점을 부여를 하였다. 이 때 주제관련 전문가 합의 방식과 응답자 평균 방식에서는 가장 높은 혹은 가장 낮은 점수를 각각 최선 혹은 최악을 반응으로

설정을 하였고, 경험적 방식에서는 준거와 가장 높은/낮은 상관을 보이는 반응을 최선/최악을 반응으로 설정하였다.

B-W(Best-Worst) 방식. 응답자가 선택한 최선과 최악의 반응에 각 채점용 답 결정방식에서 설정된 행동 반응 대안들 각각에 부여된 효과성 점수를 적용하여, 응답자가 선택한 최선의 반응에 부여되어 있는 채점용 답 점수에서 최악의 반응에 부여된 채점용 답 점수를 빼는 방식으로 점수를 산출하였다. 이 때 주제관련 전문가 합의 방식과 응답자 평균 방식에서는 각각 합의과정으로 통해 부여된 점수와 평균 점수를 각각 사용하였으며, 경험적 방식에서는 준거와 상관을 바탕으로 설정한 반응 대안의 순위에 따라 점수를 10에서 2점 단위로 순차적으로 부여를 하여 사용하였다.

최종적으로 검사의 점수는 3개의 채점용 답 결정 방식과 2가지의 채점 방식의 조합에 따라 3×2의 점수가 산출되었으며. 각 방식의 조합에 따른 신뢰도 계수는 주제관련 전문가 합의 방식에서 B-W 방식이 .65, -2~2 방식에서 .54인 것으로 나타났다. 응답자 평균 방식에서는 각각 .69와 .50으로 나타났다. 하지만 경험적 결정 방식에서는 각각 .04와 .04로 매우 낮은 문항의 내적 일관성을 나타내었다.

준거

본 연구에서는 각 응답자의 수행에 대한 기업의 성과 평가 자료를 준거로 사용을 하였다. 해당 기업의 과장급 이상의 관리직에 대해서는 모든 직원에게 동일한 척도로 평가되는 전반적인 수행(overall performance)과 리더 관리능력에 대한 수행(leadership performance)에 대한 평가를 구분하여 실시하고 있으며 평가는 해당 종업원의 직속 상사가 5점 척도를 사용하여 평가하였다. 본 연구에서 사용된 전반적 수행과 관리능력 준거 점수는 기업에서 정하고 있는 세부적 가치 행동 지표에 대한 평가점수의 평균 점수로 최종 산출, 인사 평가의 근거로 사용된 해당 기업의 인사자료를 그대로 사용하였으며, 추가적인 준거 자료로 전반적 수행 평가점수를 바탕으로 한 응답자 개인의 해당 사업부에서의 상대적인 위치를 백분위로 산출한 백분위 점수를 사용하였다. 이러한 백분위 자료를 추가적으로 사용한 것은 기업에서 사용하고 있는 인사자료의 특성상 점수의 범위와 변산이 크지 않으며, 종업원의 해당 사업부에 위치에 대한 정보를 통해 수행에 대한 추가적인 정보를 얻기 위해 사용되었다.

결 과

본 연구에서는 3가지 채점용 답 결정방식과 2가지 채점 방식에 따라 총 6가지의 상황판단검사 점수를 산출하여[1] 각각 전반적 수행에 대한 평가 점수(overall performance score: OPS)와 백분위 점수, 그리고 리더 관리 능력에 대한 평가 점수(leadership performance score: LPS)와의 상관분석을 실시하였다.

---

1) 앞서 제시한 바와 같이 경험적 방식을 사용하여 점수를 산출하였을 때에는 검사의 신뢰도가 전혀 확보되지 않는 것으로 나타났기 때문에 준거 타당도 측면에 대한 논의를 진행하는 것에 무리가 있을 것이다. 하지만 탐색적 접근의 연구라는 점에서 어떠한 결과가 산출되는지 확인해보는 것도 의미가 있다고 판단되어 경험적 방식도 준거 타당도 분석에 포함하여 실시하였다.

표 1에는 주제관련 전문가 합의 점수, 응답자 평균 점수, 준거관련 관계성 순위를 채점용 답으로 활용하여 -2~2 채점 방식을 적용한 검사 점수의 기술 통계와 준거관련 타당도가 제시되어 있다.

결과를 보면, 주제관련 전문가 방식과 응답자 방식은 각각 31.58과 32.75로 매우 유사한 평균점수와 편차를 나타낸 반면, 준거 관계성의 순위를 바탕으로 경험적으로 설정한 최선과 최악에 대한 채점용 답을 활용한 경우에는 평균이 -1.56으로 매우 낮은 평균을 나타내면서 다른 방식과 큰 차이를 보였다.

준거 타당도를 살펴보면, 주제관련 전문가 합의 점수의 경우, OPS, 백분위, LPS와 각각 .263, .226, .253으로 모두 p=.001 수준에서 유의한 것으로 나타났으며, 응답자 평균의 경우 OPS, 백분위, LPS와 각각 .266, .230, .262로 역시 모두 p=.001 수준에서 유의한 것으로 나타났다. 반면, 준거 관계성의 순위를 바탕으로 경험적으로 설정한 최선과 최악에 대한 채점용 답을 활용한 경우, LPS와는 .130으로 p=.05 수준에서 유의한 것으로 나타났으나, OPS와 백분위와는 모두 상관이 유의하지 않는 것으로 나타났다. 이를 볼 때 -2~2 채점 방식을 동일하게 적용을 할 경우, 주제관련 전문가 합의 점수와 응답자 평균을 활용하는 두 가지 방식 모두 거의 유사한 준거 관련 타당도를 나타내고 있으며, 상대적으로 응답자 평균 방식이 보다 높은 준거 타당도를 갖는 것으로 나타났다.

그 다음으로 세 가지 채점용 답 결정 방식을 활용하여 B-W 방식으로 산출한 검사 점수의 기술 통계와 준거 관련 타당도가 표 2에 제시되어 있다.

결과를 보면, 주제관련 전문가 방식과 응답자 방식은 각각 114.10과 113.94로 매우 유사한 평균점수를 나타냈으며, 편차의 경우 각각 17.71과 15.03으로 주제관련 전문가 방식에서 보다 다소 크게 나타났다. 준거 관계성의 순위를 바탕으로 경험적으로 설정한 채점용 답을 활용한 경우에는 -2~2방식에서와 마찬가지로 4.84로 상대적으로 매우 낮은 평균을 나타내면서 다른 방식과 큰 차이를 보였다. 준거 타당도를 살펴보면, 주제관련 전문가 합의 점

표 1. 세 가지 채점용 답 결정 방식과 -2~2 채점 방식 조합의 준거 관련 타당도

	평균	편차	OPS	백분위	LPS
SME X -2~2	31.58	6.19	.263***	.226***	.253***
응답자 X -2~2	32.75	5.63	.266***	.230***	.262***
경험적 X -2~2	-1.56	4.69	.055	.081	.130*

주. SME × -2~2: SME 합의점수를 채점용 답으로 하여 -2~2 채점 방식을 사용, 응답자 × -2~2: 응답자 평균점수를 채점용 답으로 하여 -2~2 채점 방식을 사용, 경험적 X -2~2: 준거 관련 순위를 채점용 답으로 하여 -2~2 채점 방식을 사용, OPS: 전반적 수행 평가 점수, 백분위: OPS으로 산출한 응답자 수행에 대한 백분위 점수로 100에 가까울수록 상위의 수행자임, LPS: 관리 능력에 대한 수행 평가 점수, 평균과 편차는 채점용 답 결정 방식과 채점 방식의 조합에 따라 산출된 총점의 평균, 편차이며, 상관계수는 각 준거와의 상관계수를 나타냄.

***p< .001

표 2. 세 가지 채점용 답 결정 방식과 B-W 채점 방식 조합의 준거 관련 타당도

	평균	편차	OPS	백분위	LPS
SME X B-W	114.10	17.71	.341***	.307***	.328***
응답자 X B-W	113.94	15.03	.289***	.265***	.288***
경험적 X B-W	4.84	20.23	.131*	.080	.159**

주. SME × B-W: SME 합의점수를 채점용 답으로 하여 최선과 최악 반응의 차이 채점 방식을 사용, 응답자 × B-W: 응답자 평균점수를 채점용 답으로 하여 최선과 최악 반응의 차이 채점 방식을 사용, 경험적 × B-W: 준거 관련 순위를 채점용 답으로 하여 최선과 최악 반응의 차이 채점 방식을 사용, OPS: 전반적 수행 평가 점수, 백분위: OPS으로 산출한 응답자 수행에 대한 백분위 점수로 100에 가까울수록 상위의 수행자임, LPS: 관리 능력에 대한 수행 평가 점수, 평균과 편차는 채점용 답 결정 방식과 채점 방식의 조합에 따라 산출된 총점의 평균, 편차이며, 상관계수는 각 준거와의 상관계수를 나타냄.

***p< .001

수는 OPS, 백분위, LPS와 각각 .341, .307, .328로 모두 p=.001 수준에서 유의한 것으로 나타났으며, 응답자 평균의 경우 OPS, 백분위, LPS와 각각 .298, .265 .288로 역시 모두 p=.001 수준에서 유의한 것으로 나타났다. 반면, 경험적으로 설정한 채점용 답을 활용한 경우, OPS와 LPS와 각각 .131, .159로 p=.05와 p=.01 수준에서 유의한 것으로 나타났으나, 백분위와는 유의하지 않는 것으로 나타났다. 이를 볼 때 B-W 채점 방식을 동일하게 적용을 할 경우, 주제관련 전문가 합의 점수를 활용하는 방식이 가장 좋은 준거 관련 타당도를 보이는 것으로 나타났다.

채점 방식간의 준거 관련성을 비교해 보았을 때, 모든 채점용 답 결정 방식에서 -2~2 방식보다는 B-W 채점 방식이 보다 높은 준거 타당도를 보였으며, 6가지의 조합 점수들 중에서는 SME × B-W의 조합이 가장 높은 준거 관련 타당도를 보이는 것으로 나타났다.

6가지 검사 점수간의 상관 분석을 한 결과가 표 3에 제시가 되어 있는데, SME × -2~2는 .881(p<.001) 로 SME × B-W와 가장 높은 관계성을 나타냈으며, 응답자 × -2~2와 경험적 × -2~2 역시 각각 .866(p<.001)과 .268

표 3. 6개 조합 방식으로 산출한 점수 간 상관

	SME X -2~2	SME X B-W	응답자 X -2~2	응답자 X B-W	경험적 X -2~2
SME X B-W	.881***				
응답자 X -2~2	.685***	.770***			
응답자 X B-W	.741***	.866***	.866***		
경험적 X -2~2	.119*	.081	.083	.047	
경험적 X B-W	.224***	.230***	.100*	.122*	.268***

주. *p< .05, ***p< .001

(p<.001)로 응답자 × B-W와 경험적 × B-W와 각각 가장 높은 관계성을 나타내었다. 서로 다른 채점용 답을 활용한 점수들과의 관계성을 살펴보면, SME × -2~2는 .741(p<.001)로 응답자 × B-W와 가장 높은 관계성을 나타내었으며, SME × B-W는 .866으로 응답자 × B-W와 가장 높은 관계성을 나타내었다. 경험적 × -2~2는 .119(p<.05)로 SME X -2~2와, 경험적 × B-W는 SME × B-W와 .230(p<.001)로 가장 높은 관계성을 나타내었다.

논 의

본 연구는 상황판단검사의 채점용 답 결정 방식과 채점 방식의 조합에 따라 검사 점수의 준거 관련 타당도가 어떠한 양상으로 나타나는지를 지식형 리더십 상황판단검사를 통해 탐색적으로 접근하여 확인하였다. 그 결과, -2~2 채점 방식을 사용할 경우, 응답자 표본의 평균 점수를 토대로 최선과 최악의 행동 반응을 선정하는 방식으로 채점용 답을 결정하였을 때 가장 높은 준거 관련 타당도를 보이는 것으로 나타났다. 이러한 준거 관련 타당도 계수는 세 가지 준거 모두에서 동일하게 나타났다. 준거 관련 타당도 계수가 응답자 표본 평균을 사용한 경우가 상대적으로 가장 크게 나타나기는 하였으나, 주제관련 전문가 집단의 합의 점수를 활용한 채점용 답 결정 방식과 비교를 할 때 그 차이는 크지 않은 것으로 나타났다. 반면, 경험적 방식으로 채점용 답을 결정하는 방식에서는 전반적인 수행에 대한 평가 점수와 개인의 상대적인 위치를 나타내는 백분위 점수와 유의하지 않은 준거 관련 타당도 계수가 관찰되었으며, 채점용 답 결정의 준거로 사용한 LPS와도 다른 방식들보다 현저하게 낮은 타당도 계수를 보이는 것으로 나타났다. 이를 바탕으로 볼 때, -2~2 채점방식을 사용하게 될 경우, 주제관련 전문가 합의 점수를 활용하거나 응답자 표본 평균 점수를 사용할 경우에는 비슷한 수준의 준거 타당도를 확보할 수 있지만, 경험적인 방식을 사용한다면 그렇지 못하다는 결론을 내릴 수 있다.

B-W 채점 방식을 사용할 경우에는 주제관련 전문가 합의 점수를 사용하는 경우가 세 가지 준거 모두와 가장 높은 준거 관련 타당도 계수를 보이는 것으로 나타났다. -2~2 방식에서 주제관련 전문가와 응답자 표본의 평균을 사용한 채점용 답에서 서로 매우 유사한 수준의 계수가 나타난 것에 비해 B-W 방식에서는 .040~.043 정도의 상대적으로 큰 차이를 나타내고 있다. 또한 -2~2 방식과 마찬가지로 경험적인 채점용 답 결정 방식이 다른 두 방식 보다 모든 준거들과 .080~.159로 상대적으로 낮은 준거 과련 타당도를 보이는 것으로 나타났다. 이러한 결과가 나타난 데는 다른 방식에 비해 경험적 채점용 답 결정 방식의 경우, 각 문항의 선택 가능 행동 대안들의 내용적인 측면을 고려하지 않고 준거와의 연관성만을 기준으로 선택 답안의 효과성 정도를 결정하기 때문에, 내용적 질의 측면 정보가 고려되지 않아 타당한 채점용 답이 산출되지 않았을 가능성을 제기할 수 있다. 이는 본 연구에서 사용한 채점용 답 결정 방식과 채점 방식의 조합에 따른 검사 문항들의 내적 일관성 계수가 전기 자료 방식을 사용하였을 경우 동일한 문항을 사용하였음에도 불구하고 상대적으로 너무 낮은 수준인 것으로 나타났다는 부분을 함께 고려할 때 가능한 설명이라 할

수 있다. 경험적 방식의 신뢰도 부분은 본 연구에서 경험적 채점용 답안 산출에 사용된 표본의 수가 크지 않은 상태에서, 상관계수에 준해 강제적으로 채점 정답 점수를 부여하는 방식을 사용하였기 때문에, 이와는 상이한 표본에 해당 채점용 답을 적용하게 되었을 때, 신뢰도를 확보하는 것이 어려웠을 수 있다. 이러한 측면에서 볼 때, 동일한 검사 문항에 상이한 채점용 답을 적용하게 되었을 경우, 이처럼 검사의 신뢰도가 확보되지 않고 결과적으로 타당도 또한 떨어지는 검사 점수가 산출될 수 있다는 것이 탐색적인 본 연구에서 확인되었다고 할 것이다.

동일한 채점용 답을 사용하고 상이한 채점 방식을 사용하였을 경우에는 모든 점수에서 B-W 채점방식이 일관되게 높은 준거 관련 타당도를 보이는 것으로 나타났으며(예를 들어, 전반적 수행 평가 준거에 대한 SME × -2~2의 .263과 SME × B-W의 .341을 비교하면, .001 수준에서 유의하게 B-W방식이 높은 준거 타당도를 보임), 이는 Knapp 등(2001)의 연구 결과와 일치하는 것이다. 이러한 결과는 -2~2 방식이 최선과 최악을 반응을 응답자가 올바르게 선택하였는지 여부에 관한 정보만을 제공하는 반면, B-W 방식은 개인의 선택한 최선과 최악의 반응 각각의 효과성 점수를 모두 고려하여 채점용 답에서 최선과 최악의 반응만이 아닌 차선책일 수 있는 반응들에 대한 응답자 선택의 정보까지 모두 제공할 수 있다는 점에서 개인 상황판단능력 정보를 보다 많이 제공할 수 있기 때문일 것이다. 이를 -2~2 방식에서는 응답자 표본 평균을 사용하거나 주제관련 전문가 합의 점수를 사용하는 것이 거의 동일한 준거 관련성 수준을 갖지만 B-W 방식에서는 주제관련 전문가 방식이 보다 높은 것으로 나타나는 것(예를 들어, 전반적 수행 평가 준거에 대한 SME × B-W의 .341과 응답자 × B-W의 .289을 비교하면, .001 수준에서 유의하게 SME 방식이이 높은 준거 타당도를 보임)에 적용을 해보면, 결과적으로 최선과 최악의 경우만을 고려하여 점수를 산출할 경우에는 주제관련 전문가들의 평가든 전문가 집단이 아닌 응답자 표본 평균을 사용하든 거의 동일한 수준의 정보를 제공해주게 되지만 B-W를 사용하게 되면 -2~2 방식보다 개인에 대한 정보를 보다 많이 제공해 줄 수 있게 되면서 채점용 답 결정 방식에 따른 정보 제공의 정도가 차이를 보이게 되는 것이라고 해석을 해 볼 수 있다. 따라서 이를 바탕으로 볼 때, 주제관련 전문가, 응답자, 상관을 사용하는 채점용 답 결정 방식 중에서 주제관련 전문가 합의 점수를 사용하는 것이 가장 준거에 대한 많은 정보를 제공해 줄 수 있을 것이며, 이 때 채점 방식은 -2~2 방식보다는 B-W 방식을 사용하는 것이 준거 타당도 측면에서는 보다 적절할 것이라는 결론을 내릴 수 있다.

이 외에 6가지 조합 점수들 간의 상관 패턴을 보면, 동일한 채점용 답을 사용한 점수들과 일관되게 가장 높은 관계성을 보이는 것으로 나타났다. 반면 동일한 채점 방식을 사용하는 것은 그러한 패턴을 보이지 않는 것으로 나타났다. 이를 볼 때, 검사 점수의 평균과 편차에는 채점 방식에 따라 차이가 날 수는 있지만, 6가지 점수의 상관분석을 토대로 해 볼 때 채점 방식보다는 채점용 답을 무엇으로 사용하는가가 개인의 검사 점수에 따른 상대적인 위치가 직접적으로 영향 받을 수 있다는 것을 알 수 있다. 이러한 결과는 상황판단검사를 개발하고 이를 시행하는데 있어 그 동안 상대적으로 관심과 그 중요성이 간과되어 온

채점용 답을 어떻게 결정하고 무엇으로 사용할 것인가의 문제가 검사의 측정적 특성과 유용성, 예측력 측면에서 중요한 이슈로 작용할 수 있음을 암시하는 부분이라 할 수 있다.

하지만 본 연구가 탐색적인 접근을 하고 있으며, 해당 기업의 인사평가 자료로 사용한 준거점수의 타당성의 확보여부의 확인이 여의치 않았다는 측면에서 연구 결과를 일반화하는데 한계점을 가지고 있다. 따라서 본 연구의 결과가 보다 일반화되기 위해서는 향후 보다 정교화 된 연구 모형을 설정하여 검증하는 추가적인 연구가 필요할 것이며, 기업의 인사평가 자료와 더불어 보다 다각화된 준거 점수의 확보가 향후 연구에 이루어져야 할 것이다. 또한 상황판단검사의 지시문 형태를 지식형으로만 사용하였기 때문에 향후에는 행동 경향형의 지시문에서도 동일한 패턴의 결과가 도출되는지 여부를 확인할 필요성이 있으며, 이와 더불어 검사가 측정하고자 하는 구성개념이 무엇인가에 따라 적합한 채점용 답안과 채점 방식의 조합이 상이할 것인가를 살펴봐야 할 것이다. 가령, 리더십의 경우, 문제 상황에서 리더가 취할 수 있는 행동 대안들이 긍정적인 결과를 도출할 수 있을 것인가에 대한 개인의 판단능력과 인지적 측면을 많이 포함되어 전문가 집단의 합의가 보다 적합할 수 있지만, 맥락수행과 같은 구성개념의 경우에는 올바른 결정인가가 아니라 얼마나 도움 행동을 적절하게 보이는가의 측면에서 접근하는 것이기 때문에 응답자 평균을 사용한 방식이 더 적합할 수도 있을 것이다.

개인의 상황판단능력에 대한 파악의 필요성과 관심이 기업이나 연구자 모두에게 증가하고 있는 상황에서 앞으로 상황판단검사는 보다 다양한 구성개념을 표적으로 더욱 다양한 검사들이 개발될 것이다. 뿐만 아니라 현재 사용되고 있는 기업의 인성 및 조직문화 적합도 검사와 같은 다양한 검사들이 지원자들의 특성을 보다 정확하게 파악하기 위해 기존과는 상이한 접근의 검사로의 전환이 요구되고 있다는 측면에서 앞으로는 상황판단형의 다양한 검사들이 개발될 가능성이 높아 질 수 있다. 이러한 맥락에서 볼 때 앞으로 상황판단검사 특성 자체에 영향을 미칠 수 있는 다양한 개발 과정의 요소를 파악하고자 하는 노력들은 끊임없이 이루어져야 할 것이다.

참고문헌

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

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

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

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. Borman(Eds.), *Personnel selection in organizations*(pp.71-98). New York: Jossey-Bass.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied*

Psychology, 76, 863-872.

Bruce, M. M., & Learner, D. B. (1958). A supervisory practices test. *Personnel Psychology, 11, 207*-216.

Cardall, A. J. (1942). *Preliminary manual for the Test of Practical Judgment.* Chicago: Science Research.

Chan**,** D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143-159.

Chan, D., & Schmitt, N. (2002). Situational Judgment and Job Performance. *Human Performance, 15(3),* 233-254

Chan, D., & Schmitt, N. (2005). Situational 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., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86(3),* 410-417.

Conn, S. R., & Rieke, M. L. (1994). The 16PF fifth edition technical manual. Champaign, IL: Institute for personality and ability Testing

File, Q. W. (1945). The measurement of supervisory quality in industry. *Journal of Applied Psychology, 29,* 381-387.

File, Q. W., & Remmers, H. H. (1948). *How Supervise? manual 1948 revision.* New York: Psychological Corporation.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-358

In J. P. Campbell & D. J. Knapp(Eds.), *Exploring the limits in personnel selection and classfication.* Mahwah, N. J.: Lawrence Erlbaum Associates.

Greenberg, S. H. (1963). *Supervisory Judgment Test manual.* Washington, DC: U. S. Civil Service Commission.

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. (1989). The reliability, validity, and practicality of the situational interview. In R. W. Eder G. R. Ferris(Eds.), The employment interview: theory, research, and practice(pp. 169-182). Newbury Park, CA: Sage.

McDaniel, M. A., Hartman, N. S., & Grubb Ⅲ, W. L. (2003, April). *Situational Judgment Tests, Knowledge, Behavioral Tendency, and Validity: A Meta-Analysis.* Paper presented at the 18th Annual Conference of the Society for Industrial and Oraganizational Psychology, Orlando, FL.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb Ⅲ, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology, 60,* 63-91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to

predict job performance: a clarification of the literature. *Journal of Applied Psychology. 86(4),* 730-740.

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

McDaniel, M. A., Yost, A. P., Ludwick, M. H., Hense, R. L., & Hartman, N. S. (2004, April). *Incremental validity of a situational judgement test.* Paper presented at the 19th Annual Conference of the Society for Industrial and Oraganizational Psychology, Chicago.

Motowidlo, S. J. (2000). Some basic issues related to contextual performance and organizational citizenship behavior in human resource management. *Human Resource Management Review, 10,* 115-126.

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

Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology. 79(4),* 475-480.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative seleciton procedure: the low-fidelity situation. *Journal of Applied Psychology, 75(6),* 640-647.

Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D. L. Whetzel & G. R. Wheaton(Eds.), *Applied measurement methods in industrial psychology.* Palo Alto, CA: Davies-Black Publishing.

Motowidlo, S.. J., Borman, W., & Schmit, M. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10,* 71-83.

Oswald, F. L., Schmitt, N., Kim. B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgement inventory as predictors of college student performance. *Journal of Applied Psychology, 89(2),* 187-207.

Parker, C.W., Golden III, J. H., Russell, D.P. & Redmond, M. R. (2000). The development of a construct-related scoring key of a situational judgment inventory for enhancing criterion-related validity. Paper presented at the 15th annual conference of the Society of Industrial and Organizational Psychology, New Orleans, April.

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),* 1-16.

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., Sydell, E. J., Scott, D. R., & Winter, J. L. (2000, April). *Factors affecting situational judgment test characteristics.* Paper presented at the 15th Annual Conference of

the Society for Industrial and Organizational Psychology. New Orleans, LA.

Smith, C. A., Organ, D. W., & Near, J. P. (1993). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology, 68,* 653-663.

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, Dallas, TX.

Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology, 52,* 1236-1247

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: the role of tacit knowledge. *Journal of Personality and Social Psychology, 49(2),* 436-458

Waugh, G. (2002). *Selecting response options and items for a situational judgment test.* Paper presented as part of the following symposium - Understanding and Predicting Performance in Future Jobs. 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50,* 25-49.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52,* 679-700.

1차 원고접수 : 2008. 10. 08
2차 원고접수 : 2009. 01. 08
3차 원고접수 : 2009. 02. 10
최종게재결정 : 2009. 02. 18

한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2009. Vol. 22, No. 1, 53-70.

The Study of Criterion-related Validities of Different Combinations of Keying and Scoring Methods in Situational Judgment Test with Knowledge Instructions

Hyun-Sun Chung **Dong-Gun Park**

Korea University

The purpose of this study is to examine the different criterion-related validities from six different combinations of three keying methods(SME consensus, average in response and empirical keying) and two scoring methods(-2~2, B-W) using a leadership situational judgment test with knowledge instructions. The test was administered to 395 employees who has managerial positions in a Korean company after developing items and deciding on keys. The results reveal that the combination of SME × B-W has the highest criterion-related validity among the six combinations. This study suggests that the keying method of a situational judgement test is one of the important factors that determine the psychometric and predictive aspects of the test.

Key words : Situational Judgment Tests, Keying, Scoring, Leadership