한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2025. Vol. 38, No. 3, 313-334
https://doi.org/10.24230/kjiop.v38i3.313-334

# 업무 시뮬레이션을 활용한 상호평가 기반 논술형 SJT와 직무적성검사의 타당도 비교*

김 송 의　　박 주 용†
서울대학교 심리학과

조직에서 우수 인재 확보하려면 선발 도구의 예측 타당도가 높아야 한다. 본 연구는 실제 직무 상황을 반영한 업무 시뮬레이션 과제 수행을 준거로, 논술형 상황판단검사(Situation Judgment Task: SJT)와 구성형 및 선다형 직무적성검사의 상대적 예측력을 비교하였다. 이를 위해 분석적 사고력, 의사소통능력, 문제해결능력을 평가하도록 설계된 네 가지 시뮬레이션 과제를 개발하였다. 대학교 재학생 및 졸업생 50명을 대상으로 논술형 SJT, 직무적성검사, 그리고 개발된 시뮬레이션 과제를 수행하도록 하고, 업무 시뮬레이션 과제 수행 점수를 종속변수로 설정하여 SJT 점수와 직무적성검사 점수를 독립변수로 회귀분석을 수행하였다. 이와 함께, 참여자 간 상호평가를 통해 채점한 SJT 결과와 전문가 평정 결과 간의 일치도를 통해 상호평가 채점 방식의 준거타당도를 검토하였다. 그 결과, 논술형 SJT 점수는 의사소통 능력과 문제해결 능력 과제 수행을 유의미하게 예측한 반면, 직무적성검사 점수의 예측력은 모든 영역에서 미미한 것으로 나타났다. 또한, 상호평가를 기반으로 한 채점 결과와 전문가 평정 간에는 유의미하게 높은 상관이 관찰되었다. 이상의 결과가 갖는 이론적, 실무적 함의와 후속 연구 방향에 대해 논의하였다.

주요어 : 상황판단검사, 상호평가, 직무적성검사, 업무 시뮬레이션, 타당도, 인재 선발

---

* 세심하고 아낌없는 조언을 해주신 편집위원과 익명의 심사위원들께 감사를 드립니다.
본 논문은 김송의의 서울대학교 박사학위논문인 “조직의 선발 과정에서 구성형 문항과 상호평가를 이용한 논술의 타당도 검증”을 발전시킨 결과물임을 밝힙니다.

† 교신저자 : 박주용, 서울대학교 심리학과 교수, 서울특별시 관악로1 서울대학교
Tel: 02-880-9050, E-mail: jooypark@snu.ac.kr

Copyright ⓒ 2025, The Korean Society for Industrial and Organizational Psychology. This is an open-access article distributed under the terms of the Creative Commons Attribution –NonCommercial Licenses (https://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

## 서 론

우수 인재 확보는 조직의 성과와 경쟁력을 좌우하는 핵심 요소이다. 특히 정보통신 산업에서는 우수 인재가 일반 직원 대비 수백 배 이상의 가치를 창출한다고 평가된다(Tam & Delaney, 2005). 같은 맥락에서 구글(Google)의 인사 담당자인 라즐로 복(Laszlo Bock)은 "교육과 훈련보다 채용에 투자하라"고 강조한 바 있다(Bock, 2015). 이러한 흐름은 국내에서도 예외가 아니다. 최근 국내에서는 “한 명의 천재가 10만 명을 먹여 살린다”는 통념이 확산될 정도로 우수 인재의 중요성에 대한 인식이 높아졌으며, 실제 주요 기업들은 급변하는 글로벌 환경 속에서 핵심 인재의 확보와 유지를 위해 막대한 노력을 기울이고 있다(Shvetsova, 2025).

선발은 단순히 높은 시험 점수를 가진 지원자를 선별하는 과정이 아니다. 그보다는 실제 직무 상황에서 탁월한 성과를 발휘할 수 있는 역량을 지닌 인재를 가려내는 과정이다. 이를 위해 신뢰도와 타당도가 확보된 선발 도구의 사용이 필수적이며, 지원자들의 반응, 도구의 부작용, 비용과 시간 등 다양한 요인을 고려해야 한다(Schmidt & Hunter, 1979, 1998; Van Iddekinge et al., 2023). 그러나 많은 기업이 여전히 객관식 중심의 정형화된 평가 방식에 의존하고 있다(이승철 & 어수봉, 2017). 이 방식은 신뢰도 확보에는 유리하나 실제 직무 역량이나 복합적인 문제해결 과정을 충분히 반영하기 어렵다는 한계가 있다(조성준 & 박찬균, 2020).

이런 문제를 보완하기 위해 상황판단검사(Situational Judgment Test; SJT)가 활용되어 왔다. SJT는 실제 직무 상황과 유사한 시나리오를 제시하여 지원자의 맥락적 사고와 문제 해결 과정을 직접 관찰할 수 있고, 높은 타당도를 보이는 것으로 알려져 있다(McDaniel et al., 2001; McDaniel et al., 2007). 기존에는 주로 선택형 문항으로 제시되었으나, 최근에는 게이미피케이션 요소가 결합되거나(Landers & Sanchez, 2022), 구성형 응답 방식을 통해 지원자의 사고 과정을 심층적으로 평가하는 시도가 이루어지고 있다(Iqbal et al., 2025). 특히 이러한 구성형 SJT는 단편적 지식보다는 종합적・고차원적 사고 능력을 파악하는 데 효과적이어서, 의대 선발이나 교육・훈련 프로그램에서도 점차 그 활용 범위를 넓혀가고 있다(Mortaz Hejri et al., 2023).

그렇지만 논술형 SJT는 채점에 소요되는 시간과 인적 자원, 평가자 간 일관성 확보 등의 실무적 부담이 크다는 결정적인 약점이 있다. 교육 현장에서는 이러한 약점을 극복하기 위해 상호평가(peer assessment)가 널리 사용된다. 상호평가는 응시자가 서로의 답안을 평가하도록 하는 방식으로, 교육적 효과뿐 아니라 전문가 수준의 평가 정확도를 확보할 수 있다는 연구 결과가 많다(Falchikov & Goldfinch, 2000; Li et al., 2016; 박정애 & 박주용, 2019; Liu & Carless, 2006; Schunn et al., 2016). 이런 결과는 논술과 같은 구성형 응답을 상호평가로 채점하면, 채점 부담은 줄이면서도 깊이 있는 사고 과정을 평가할 수 있는 대안이 될 수 있음을 시사한다.

본 연구는 실제 직무수행 상황을 반영한 업무 시뮬레이션 과제를 준거로 삼아, 논술형 SJT와 직무적성검사의 예측 타당도를 비교하고자 하였다. 구체적으로 분석적 사고력, 문제해결능력, 의사소통능력 세 영역을 측정하는 네 가지 업무 시뮬레이션 과제를 설계하고,

논술형 SJT와 직무적성검사를 시행한 후 회귀분석으로 수행 결과에 대한 예측력을 비교하였다. 또한, SJT 답안에 대해 상호평가 방식의 채점 결과와 전문가 채점 결과의 일치도를 분석하여 상호평가 기반 채점 방식의 준거 타당도를 검토하였다. 이 과정을 통해 상호평가 기반 논술형 SJT가 실무적 부담을 줄이면서도 타당도 높은 선발 도구로 활용될 수 있는지를 확인하고자 한다.

## 이론적 배경 및 가설 설정

### 일반인지능력(GMA)과 직무적성검사

일반인지능력(GMA; general mental ability)은 언어 이해, 수리적 문제해결, 추론 등 여러 인지 영역에 공통적으로 기여하는 지적 능력으로, 전통적으로 Spearman(1904)이 제안한 *g* 요인으로 정의된다(Carroll, 1993). 메타분석 결과, GMA는 직무 경험이 없는 지원자의 직무성과를 예측하는 가장 강력한 요인이며(Schmidt & Hunter, 1998; Gottfredson, 1997), 예측타당도 계수는 약 .40~.62 수준에 이른다(Salgado et al., 2003; Bertua et al., 2005; Salgado & Moscoso, 2019). 특히, 직무 복잡도가 높아질수록 GMA의 예측 타당도가 향상되며(Salgado & Moscoso, 2019), 이러한 예측력은 산업과 직무를 초월하여 일관되게 나타난다(Hunter & Hunter, 1984). 또한 GMA 측정과 구조화된 면접을 병행할 경우 선발 타당도가 더욱 향상될 수 있는데($r$ = .63; Schmidt & Hunter, 1998), 국내에서도 표준화된 질문과 평가 기준에 기반한 구조화 면접 절차의 필요성이 제안된 바 있다(오인수 & 서용원, 2002).

일반인지능력(General Mental Ability; GMA)은 다양한 인지 하위능력들에 공통적으로 존재하는 요인으로, 구체적으로는 어휘 이해, 독해, 논리적 언어 추론 등으로 구성된 언어 능력과, 산술 계산, 수열 추리, 자료 해석 등을 포함한 수리 능력이 핵심적이다(Carroll, 1993; Schneider & McGrew, 2012). 예컨대, 미국 노동부의 일반적성검사 배터리(General Aptitude Test Battery: GATB)는 언어, 수리, 공간 등의 개별 하위검사 점수를 통합하여 전체적인 GMA 점수를 산출하는 방식으로, 어휘력, 산술 추론 및 계산, 공간지각 과제 점수를 합산해 *g* 요인을 측정하고 있다(U.S. Department of Labor, 1981). 또한, ICAR 지능검사 개발 연구에서는 언어추론 및 문자 · 숫자 추리 문항들이 약 0.8의 높은 *g* 요인 적재량을 나타냈는데(Condon & Revelle, 2014), 이는 GMA를 대표하는 핵심 지표로서 언어 이해력과 수리 추리력이 가지는 중요성을 실증적으로 뒷받침한다.

한편, 우리나라 일반적성검사 문항은 주로 선다형이다(이승철 & 어수봉, 2017). 선다형은 지문, 발문, 보기, 선택지로 구성되어 미리 제시된 답지 중 정답을 선택하도록 하며, 출제와 채점이 용이하므로 대규모 시험에서 선호된다. 이러한 방식은 시간과 비용 측면에서 효율적이지만, 피평가자가 단순 추측으로 정답을 맞힐 가능성이 있어 실제 능력 평가에는 한계가 있다. 또한, 주어진 선택지 중 하나를 고르는 수동적 응답 방식이라는 점도 문제점으로 지적된다. 이러한 선다형 문항의 한계를 보완할 수 있는 방식이 구성형 문항이다. 구성형 문항은 피평가자가 스스로 답안을 작성하도록 요구하여 보다 깊은 사고 과정을 유도할 수 있다는 장점이 있으나, 채점 부담이 커

대규모 평가에서는 활용이 어렵다는 현실적 제약이 있다. 하지만 최근에는 컴퓨터 기반 검사(Computer-based Testing; CBT)가 보편화됨에 따라 선다형과 구성형의 장점을 결합한 새로운 응답 방식이 시도되고 있다. 예컨대, 짧은 답변을 직접 입력하도록 하여 구성형의 장점인 추측 가능성을 최소화하면서도, 컴퓨터 기반 자동 채점을 통해 선다형의 효율성 또한 일부 확보할 수 있게 되었다. 이러한 기술적 발전은 향후 일반적성검사의 형식과 평가 방식에 변화를 가져올 것으로 기대된다.

## 상황판단검사

상황판단검사(Situational Judgement Test; SJT)는 응시자에게 직무나 학업 환경에서 마주할 수 있는 가상의 시나리오를 제시한 뒤, 가장 적절하다고 판단되는 행동이나 의사결정을 선택하거나 서술하도록 요구하는 평가 방법이다(이상철 등, 2003; McDaniel et al., 2007). SJT는 선다형, 다지선다형, 순위 매기기 등의 선택형, 그리고 응시자가 자신의 언어로 답안을 기술하게 하는 구성형 방식으로 구분된다(McDaniel & Nguyen, 2001). 선택형은 고정 응답 형식으로 인해 인구통계학적 집단 간 차이를 더 크게 유발하거나 위조·기만에 취약할 수 있고(김의수 등, 2011), 채점 방식이 신뢰도와 타당도에 영향을 미칠 수 있다(정현선 & 박동건, 2009). 반면, 구성형 문항은 응답의 다양성이 높아 채점 변동성이 더 커질 수 있다는 단점이 있다(McDaniel et al., 2007). 그럼에도 불구하고 제시된 상황에서 응시자가 스스로 어떻게 행동할지 서술하게 하는 논술형 SJT는 공감 능력, 윤리적 사고, 대인관계 능력, 성실성 등 비인지적 특성을 보다 직접적으로 측정할 수 있게 한다. 이 때문에, SJT는 의과대학 입학 선발 및 훈련 현장뿐만 아니라 국내외 기업의 채용 및 평가 과정에서도 널리 활용된다(Patterson et al., 2015; Mortaz Hejri et al., 2023).

전통적 성격 검사 척도는 응답자가 자신을 긍정적으로 보이도록 응답을 왜곡하거나 위조·기만할 가능성이 크다. 이는 특히 입학이나 채용 선발 상황에서 검사 점수의 신뢰성과 타당성을 저하시킬 수 있다(Brannick & Ellingson, 2006). 반면, SJT 형식을 활용하면 이러한 왜곡을 방지할 수 있다. 예컨대 권은아와 이종구(2021)의 연구는 SJT로 측정한 정직성 척도가 실제 업무성과와는 유의미한 정적 상관을, 역 생산적 행동과는 유의미한 부적 상관을 보인다는 것을 확인하였다. 더 나아가, 논술형 SJT는 응시자가 구체적인 시나리오에 대해 직접 판단이나 행동을 서술하도록 요구하므로 응답의 진실성을 높이고 위조·기만 시도를 쉽게 발견할 수 있는 장점이 있다(Kasten et al., 2020).

## 평가센터와 업무 시뮬레이션 과제

전 세계 여러 조직에서는 지원자의 직무수행 역량을 다면적으로 평가하기 위해 평가센터(assessment center)를 활용해왔다. 평가센터는 조직이 요구하는 핵심 역량을 토대로 직무 상황을 시뮬레이션한 과제를 설계하고, 복수의 평가자가 행동 관찰 척도나 체크리스트를 통해 지원자의 행동을 기록·평가하여 종합 판단을 내리는 방식이다. 이러한 기법은 제2차 세계대전 전후 군사 분야 장교 선발에서 시작되어 민간 조직으로 확산되었다. AT&T의 장기 종단 연구를 통해 관리자 후보 예측에 효

과적이라는 점이 입증된 이후 IBM, GE, Sears 등 다수 기업으로 확산되었다(Krause & Gebert, 2003).

우리나라에서는 주로 공공부문 승진 선발에서 평가센터를 도입하여 활용 중인데, 직무 분석을 기반으로 한 과제로 역할 연기(role-play), 인바스켓(In-basket), 집단 토론(Group Discussion), 프레젠테이션, 사례분석 · 보고서 작성, 작문 과제 등 다양한 형태의 시뮬레이션 과제가 사용된다. 각 과제는 해당 직무에서 실제로 요구되는 상황과 행동 지표를 반영하도록 설계되며, 지원자가 보이는 행동을 사전에 정의된 평정 요소에 따라 독립적으로 기록하고, 평가자 간 합의를 통해 신뢰도를 확보한다. 메타분석 결과, 평가센터의 전체 평가점수와 직무 성과 간 예측타당도는 약 .37이며(Gaugler et al., 1987), 이후 연구에서 분석한 상관계수는 약 .28이었다(Hermelin et al., 2007). Arthur 등(2003)은 평가센터 6개 차원의 타당도가 .25에서 .39 범위에 있음을 보고하였는데, 이는 과제 설계 및 차원 선택에 따라 타당도가 달라질 수 있음을 시사한다.

## 공통 핵심 역량

특정 직무나 직위에 관계없이 모든 직업인에게 공통적으로 요구되는 기본적인 능력과 자질을 분석한 연구들도 있다. 예컨대 미국 교육부의 고용 역량 틀(Employability Skills Framework)에 따르면, 구직자가 갖춰야 할 자질은 응용 지식(applied knowledge), 작업 기술(workplace skills), 효과적 관계(effective relationships)의 세 가지 차원으로 구성된다. 각 차원 아래 비판적 사고, 자원관리능력, 의사소통능력, 시스템 사고(systems thinking) 등 총 9개의 세부 역량이 포함된다(미국 교육부, n.d.). 이와 유사하게 미국 비즈니스 및 산업협회(National Network of Business and Industry Associations; NNBIA, 2014)는 응용 지식, 작업 기술, 개인 기술(personal skills), 대인 기술(people skills)의 네 가지 차원을 제안했다. 이 연구에서는 읽기, 쓰기, 문제해결, 비판적 사고, 의사결정, 적응성, 의사소통능력 등 20가지의 세부 역량이 예비 경제인에게 필수적이라고 밝혔다. 또한 영국 국립 과학기술 예술기금(National Endowment for Science, Technology and the Arts; NESTA, 2017)은 수요가 높은 직업군에서 공통적으로 요구되는 기술로 고차적 인지 기술(higher-order cognitive skills), 시스템 기술(systems skills), 대인관계 기술(interpersonal skills)을 제시하였다.

우리나라의 경우 한국산업인력공단(2020)을 중심으로 모든 직업인에게 요구되는 공통 역량을 9개 영역으로 구분했다. 의사소통능력, 문제해결능력, 대인관계능력 등이 대표적이며, 문서작성능력, 비판적 사고, 갈등관리능력 등이 구체적인 하위 역량으로 제시된다.

여러 연구에서 강조된 핵심 역량을 종합하면, 공통적으로 사고력, 문제해결능력, 대인관계능력, 의사소통능력 등이다. 이 중 대인관계능력과 의사소통능력은 연구에 따라 범주 명칭은 조금씩 다르지만 대체로 유사한 개념으로 분류된다. 그러나 사고력과 문제해결능력은 연구자나 기관마다 분류 방식에 차이가 나타난다. 예컨대 미국 교육부는 사고능력을 논리력, 비판적 사고력, 문제해결능력 등을 포괄하는 상위 개념으로 설정한 반면, 한국산업인력공단(2020)은 문제해결능력 아래에 비판적 사고력과 창의적 사고력 등을 포함시키고 있다. 이처럼 사고력과 문제해결능력 간의 관계

는 아직 명확히 체계화되어 있지 않지만, 다양한 직무에서 필수적으로 요구되는 역량이 사고력, 문제해결능력, 대인관계능력, 의사소통능력임은 분명하다.

본 연구에서는 선행연구를 바탕으로, 평가센터 방식의 업무 시뮬레이션 과제를 준거 과제로 사용하였다. 이를 위해 국내외 선행연구에서 일관되게 강조된 분석적 사고, 문제해결능력, 의사소통능력을 중심 역량으로 설정하고, 4개의 세부 과제를 개발하였다. 세부 과제는 각각 분석적 사고과 문제해결능력을 묻는 과제 한 가지와 의사소통능력을 묻는 과제 두 가지다. 개발된 업무시뮬레이션 과제를 준거로, 직무적성검사와 SJT의 예측력을 비교하였다. 직무적성검사 문항으로는 배경지식 영향을 줄이면서 수리 능력과 언어능력을 각각 평가할 수 있는 자료해석 유형과 문단배열 유형을 활용하였다. SJT 문항은 조직 내 실제 문제 상황을 제시하고, 응시자가 해결 방안과 그 근거를 서술하도록 설계하였다.

SJT는 의사결정능력, 문제해결능력, 대인관계능력, 팀워크 능력 등 비인지적 역량을 신뢰도 높게 측정하는 도구로, 응시자의 다양한 상황에서의 행동에 대한 절차적 지식을 평가할 수 있기 때문에(Lievens & Sackett, 2012; Patterson et al., 2016), 전통적 직무적성검사보다 분석적 사고, 의사소통능력, 문제해결능력에 대한 예측력이 더 높을 것으로 기대할 수 있다. 따라서 네 가지 세부과제로부터의 점수를 합산한 업무 시뮬레이션 과제에서 SJT의 예측력이 직무적성검사보다 더 높을 것으로 예상하였다. 또한, 상호평가의 타당도에 대한 선행논문을 바탕으로 상호평가 기반 채점 결과와 전문가 평정 간의 상관이 유의미하게 높을 것으로 예상하였다(Falchikov & Goldfinch, 2000; Li et al., 2016). 이에, 본 연구의 가설은 다음과 같다.

**가설** 1. 직무적성검사보다 SJT가 업무 시뮬레이션 과제를 더 잘 예측할 것이다.

**가설** 1-a. 직무적성검사보다 SJT가 분석적 사고 과제를 더 잘 예측할 것이다.

**가설** 1-b. 직무적성검사보다 SJT가 의사소통능력 과제를 더 잘 예측할 것이다.

**가설** 1-c. 직무적성검사보다 SJT가 문제해결능력 과제를 더 잘 예측할 것이다.

**가설** 2. SJT의 상호평가 결과가 전문가 결과와 유의미하게 높은 일치도를 보일 것이다.

## 방 법

### 연구 대상

본 연구는 연구윤리위원회(IRB No. 2311/003-009)의 승인을 받아 진행되었으며, 연구 참여자들은 자발적으로 동의서에 서명하고 참여에 응하였다. 참여자는 S대학교 커뮤니티를 통해 5인 1조로 모집되었으며, 서울 소재 대학교 재학생 및 졸업생 총 50명이 참여하였다. 이 중 재학생은 45명, 졸업생은 5명이었고, 전공별로는 사회과학・인문대 학생 20명, 공과대・자연대 학생 18명, 그 외 사범대, 음악대, 의과대 학생들이 포함되었다.

참여자들은 논술형 SJT, 직무적성검사, 시뮬레이션 과제를 순차적으로 수행하였으며, 미완결 데이터를 포함한 3개 사례는 최종 분석에서 제외하였다. 최종 분석 대상은 47명으로, 이들의 평균 연령은 25.51세($SD$ = 3.28)였으며, 성별 분포는 남성 22명, 여성 25명이었다.

연령 및 성별에 따른 유의미한 차이는 확인되지 않았다.

## 연구 절차

본 연구는 피험자 내(within-subjects design) 설계로, 각 연구참여자는 휴식시간을 포함해 총 120분간, 3가지 과제를 수행하였다(그림 1). 과제는 SJT, 직무적성검사, 시뮬레이션 순으로 진행되었으며, SJT는 40분, 직무적성검사는 30분, 시뮬레이션 과제 풀이는 40분간 이루어졌다. 구체적으로 SJT는 최대 25분간 답안을 작성하도록 한 후, 최대 15분간 타인 답안을 상호평가 하도록 하였다.

## 자극 및 측정 도구

본 연구에서 사용된 자극은 논술형 SJT, 직무적성검사, 그리고 업무 시뮬레이션 과제였다.

### 논술형 SJT

조직 내에서 발생할 수 있는 시나리오를 읽고, 본인이라면 제시된 상황에서 어떻게 행동할지를 두 가지 해결책을 중심으로 서술하도록 하였다. 제시된 시나리오는 기업 내 360° 다면평가 프로그램을 처음 도입하는 과정에서 발생한 문제 상황으로, 프로그램의 원활한 운영을 위해 응시자가 HR 매니저로서 해결책을 모색하도록 구성하였다. 답안의 분량은 2~3문단으로 제한하였으며, 외부 검색 도구나 타인의 도움 없이 응시자가 자율적으로 작성하도록 통제하였다. 답안 작성과 익명 상호평가는 '클래스프랩(Classprep)'이라는 온라인 시스템을 통해 이루어졌다.

### 직무적성검사

본 연구에서는 6개의 자료 해석 문항(그림 2)과 6개의 문단 배열 문항(그림 3)으로 구성된 총 12개의 검사 문항을 사용하였다. 단순 추측으로 인한 배점 가능성을 배제하고자 먼저 지문과 발문, 보기를 읽고 정답을 직접 '구성형'으로 응답하도록 하였으며, 이후 선택지가 추가로 제시된 상황에서 옳은 선택지를 다시 '선택'하도록 하였다. 실험 과정에서는 응답자에게 신속하고 정확한 문제 풀이를 안내하였고, 오답 시 감점이 적용됨을 사전 고지하여, 정답 선택 시 추측 가능성을 최소화하고 응답자의 실제 능력이 반영되도록 하였다.

모든 문항 응답은 컴퓨터를 통해 이루어졌으며, 한 번 응답을 완료한 후에는 이전 문제로 돌아가 응답을 수정할 수 없도록 하였다. 이는 이후 단계에서 제시되는 선다형 선택지가 정답에 대한 단서로 작용하는 것을 방지하기 위한 조치였다. 검사는 시간이 종료되거나 모든 문항 풀이가 완료되면 자동으로 종료되도록 설계하였으며, 계산이 필요한 문항에 대비하여 펜과 종이를 별도로 제공하였다.

### 업무 시뮬레이션 과제

업무 시뮬레이션 과제 개발을 위해 전공 분

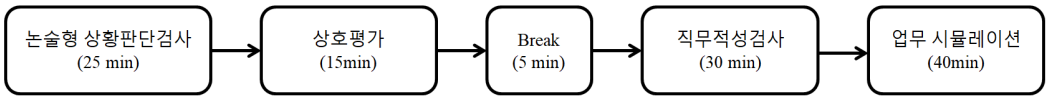



그림 1. 연구 절차




그림 2. 도표해석 예시문항

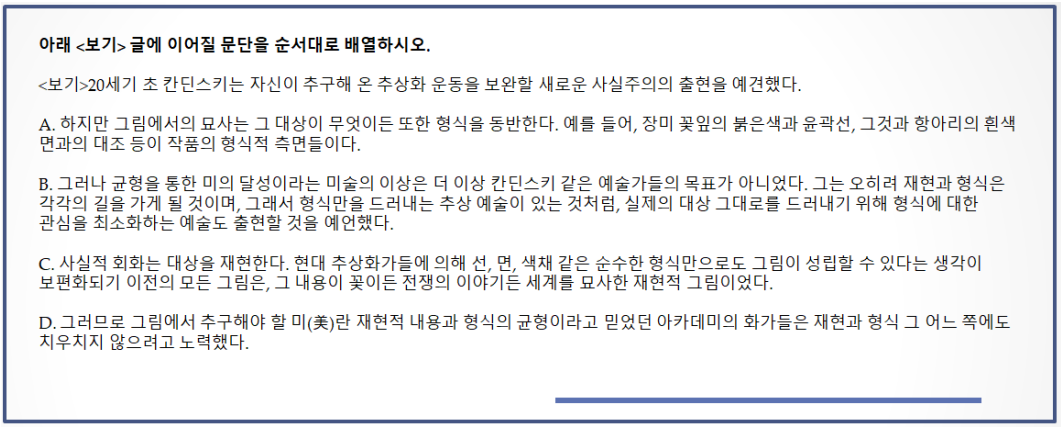



그림 3. 문단배열 예시문항

야 교수와 논의를 거쳐 내용 타당도를 확보하였다. 또한, 소규모 표본을 대상으로 예비 검사를 실시하여 각 문항의 난이도, 풀이 시간, 그리고 평가 결과의 변별력을 점검하였다. 풀이 시간을 고려한 후, 평가 목적에 가장 부합하는 네 가지 세부 과제를 최종 선정하였다. 선정된 세부 과제는 온라인 연구 홈페이지를 통해 제시되었으며, 참여자들은 주어진 상황을 읽고 제한된 시간 내에 과제 풀이를 수행하였다. 각 세부 과제는 기업 인사팀 대리가 하루 동안 수행해야 하는 네 가지 업무 상황으로 구성되었다. 참여자들은 1, 2, 3, 4의 순서에 따르지 않고, 자율적으로 원하는 순서에 따라 과제를 해결할 수 있도록 하였다.

참여자들이 수행해야 할 세부과제는 평판 조회를 위한 리스트를 작성하는 것, 인턴사원에게 불합격 메일을 보내는 것, 업무 만족도 조사 문항을 작성하는 것, 그리고 요청받은

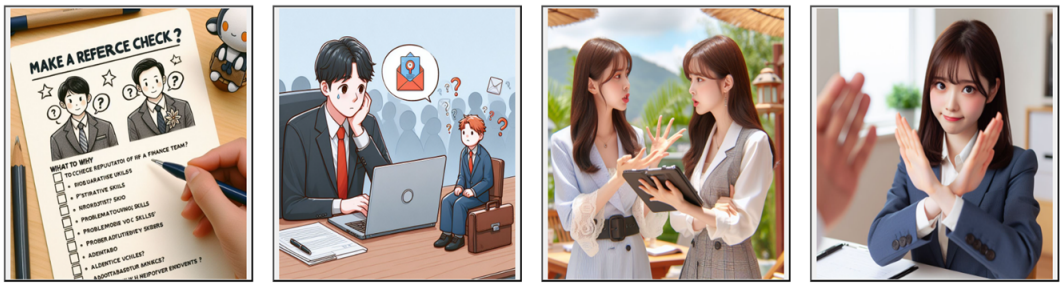



그림 4. 업무 시뮬레이션 과제 : 네 가지 세부 과제

부정 인사청탁에 대한 거절 메일을 보내는 것으로 구성되어 있었다(그림 4). 연구참여자들은 연구 홈페이지에 제시된 내용 중 과제 수행에 필요한 정보를 취사선택해야 했다. 예를 들어, 인턴사원의 메일 주소와 부정 청탁을 요청한 차장의 메일 주소를 홈페이지에서 확인하여 작성해야 했으며, 이러한 정보를 정확히 확인하고 과제를 수행했는지, 그리고 주어진 과제를 정확히 이해하고 수행하였는지 등이 각 세부 과제를 평가하는 기준이었다.

그 중 첫 번째 과제는 평판 조회 문항을 작성하는 것이다. 다양한 질문이 떠오를 수 있지만, 이를 협업 능력, 윤리의식, 문제해결능력, 적응력이라는 네 가지 범주에 맞추어 7가지로 작성하도록 하였다. 이후 각 범주에 가장 적합한 우수 문항 4가지를 선택해 매칭하도록 하였다. 이를 위해서 연구참여자들은 각 범주의 개념을 정확히 이해한 후, 관련성이 낮은 질문은 제거하는 한편, 적합한 질문은 범주에 맞게 분류하도록 했다. 따라서 여러 가지 질문을 떠올리고, 이를 세분화하고 체계적으로 분류하는 분석적 사고가 부각될 것으로 예상하였다.

두 번째 과제는 인턴사원에게 불합격 메일을 작성하는 것이다. 불합격 메일은 단순 통보가 아닌, 수신자에 대한 배려와 존중을 담아 작성되어야 한다. 따라서 상대방의 감정을 고려한 세심한 의사소통능력이 요구될 것이라 예상하였다. 마찬가지로 네 번째 과제는 요청받은 부정 인사청탁에 대한 거절 메일을 작성하는 것이었는데, 거절 메일은 단호함을 유지하면서도 상대방의 체면을 고려해 정중하게 작성해야 한다. 이를 위해 거절의 이유를 명확히 전달하면서, 부정적인 감정을 최소화하는 어조를 사용하도록 하였다. 이처럼 두 번째 과제와 네 번째 과제는 연구참여자가 불편함을 겪을 만한 어려운 상황에서 어떤 방식으로 의사소통하는지를 평가하며, 더 나아가 정서적 지능과 대인관계능력이 효과적으로 반영되도록 구성하였다.

세 번째 과제는 신입직원 만족도 문항을 작성하는 것이었는데, 연구참여자들은 신입직원들의 퇴사율이 높은 조직 내 상황을 문제 상황으로 인식하고, 그 원인을 다양하게 분석해 각 원인에 대해 만족도를 묻는 질문을 생성해야 했다. 예를 들어, 업무 환경, 상사와의 관계, 조직 문화 등 신입직원들의 퇴사를 초래할 수 있는 원인을 범주화하고, 이를 바탕으로 질문을 구성하도록 기대하였다. 따라서 이 과정에서 문제의 원인을 정확히 파악하고, 개

선 방안을 고민 및 제시하는 문제해결능력이 가장 많이 요구될 것이라 예상하였다. 각 과제는 평가 목적과 난이도, 풀이 시간을 고려해 제시되었다.

## 변수 및 측정치

본 연구에서는 논술형 SJT, 직무적성검사, 시뮬레이션 과제 수행을 측정하였다.

### 논술형 SJT

SJT는 논술형으로 제시되었으며, 독창성과 흐름의 두 가지 기준에 따라 채점되었다. 각 기준은 7점 리커트(Likert) 척도로 평가하였다. 독창성 평가는 해결 방안이 명확하고 구체적일수록 높은 점수를 부여하였으며, 흐름 평가는 아이디어 간의 인과 관계 및 우선순위 판단 근거가 논리적으로 전개되었는지에 중점을 두었다. 평가는 전문가 점수와 상호평가 점수를 각각 산출하여 이루어졌다. 전문가 점수는 박사 과정 이상 1인과, 조직 실무 경험 10년 이상인 1인의 평가자가 실시한 후 평가 일치도를 확인하였다. 평가자 간 신뢰도는 급내상관계수(Intraclass Correlation Coefficient: ICC)를 사용하여 평가되었다. 본 연구에서는 두 명 이상의 평가자가 고정된 그룹으로 점수를 매기는 ICC(3,k)를 사용하였다. ICC(3,k)는 평가자 간 일관성을 측정하며, 값이 1에 가까울수록 평가자 간의 일치도가 높은데(Shrout & Fleiss, 1979), (1) 독창성에 대한 ICC(3,k)는 .77, 95% CI [.61, .87]였으며, (2) 흐름에 대한 ICC(3,k)는 .85, 95% CI [.74, .92]였다.

이후, 온라인 상호평가 시스템에서 산출한 점수를 통해 보정된 상호평가 점수를 산출하였다. 온라인 상호평가 시스템에서 산출된 원점수에 각 평가자의 평가능력을 곱하여 가중평균(weighted mean)을 구한 뒤, 이를 70점 척도로 환산하여 보정된 상호평가 점수를 산출하였다. 이러한 보정점수는 평가자의 평가능력을 반영했을 때 산출된 상호평가 점수가 전문가 평가 점수와 어느 정도 일치하는지를 검토하기 위해 계산되었다.

### 직무적성검사

직무적성검사 점수는 각 참여자가 12개 문항 중 맞힌 문항 수에 따라 배점된 점수를 활용하였다. 선다형과 구성형, 두 가지 방식으로 응답하도록 하였는데 선다형의 경우 맞혔을 경우 10점, 틀렸을 경우 0점을 부여하는 방식으로 배점이 이루어졌다. 구성형은 도표해석 문항과 문단배열 문항의 채점 방식이 상이했는데, 도표해석 문항은 옳은 선택지를 모두 선택하도록 하고 무응답 시 0점을, 옳은 선택지를 고를 때마다 가점을 부여하였다. 반면, 틀린 선택지를 고를 때마다 감점을 적용하였다. 문단배열 문항은 기본 정렬 알고리즘 중 하나인 버블소트(bubble sort) 방식에 따라 이동이 필요한 횟수만큼 감점을 적용하였는데, 예를 들어 ABCD가 정답인 문항에서 ABDC로 답변하면 1번 이동만큼의 감점이 이루어지는 식으로 채점되었다.

### 업무 시뮬레이션 과제

업무 시뮬레이션 과제의 각 세부 과제는 11점 만점으로 두 명의 전문가가 채점한 평균값을 사용하였고, 각 과제 수행을 각각 분석적 사고, 의사소통능력, 문제해결능력 점수로 활용하였다. 분석적 사고와 문제해결능력을 묻는 세부과제는 한 가지씩이었으나 의사소통능력을 묻는 세부과제는 두 가지였는데, 의사소

통능력 점수는 두 개 세부과제의 합산 점수를 활용하였다. 이후, 네 가지 세부 과제의 점수를 합산하여 시뮬레이션 과제 점수로 사용하였다.

이후, 시뮬레이션 세부과제에 대한 평가는 박사 과정 이상 1인과, 실무 경험 10년 이상인 1인의 평가자가 실시한 후 평가 일치도를 확인하였다. 평가자 간 신뢰도는 급내상관계수(Intraclass Correlation Coefficient: ICC)를 사용하여 평가되었다. 본 연구에서는 두 명 이상의 평가자가 고정된 그룹으로 점수를 매기는 ICC(3,k)를 사용하였다. 분석적 사고에 대한 ICC(3,k)는 .95, 95% CI [.91, .97]였으며, 의사소통능력 과제(1)에 대한 ICC(3,k)는 .95, 95% CI [.91, .97]였다. 의사소통능력 과제(2)에 대한 ICC(3,k)는 .96, 95% CI [.93, .98]였으며 문제해결능력에 대한 ICC(3,k)는 .99, 95% CI [.97, .99]였다.

### 분석 방법

본 연구에서는 기술통계, 상관계수 분석 및 회귀분석을 실시하였다. 이를 위해 변수들의 평균, 표준편차를 확인한 이후, 각 변수가 업무 시뮬레이션 과제를 각각 얼마만큼 설명하는지 검증하기 위해 선형 회귀분석을 실시하였다.

## 결 과

### 기술통계 및 상관계수표

논술형 SJT의 상호평가 점수는 독창성 평균 4.16점($SD$ = 0.49), 흐름 평균 4.04점($SD$ = 0.59)이었다. 전문가 평가는 독창성 평균이 4.44점($SD$ = 1.03)으로 상대적으로 높았으며, 흐름 평균은 4.03점($SD$ = 0.63)이었다. 평가자의 평가능력을 반영하여 가중평균을 산출하고, 이를 70점 척도로 환산하여 보정된 상호평가 점수를 구한 결과, 평균은 50.04점($SD$ = 6.86)이었다.

본 연구에서 실시한 Pearson 상관분석(N = 50, df = 48) 결과, 업무 시뮬레이션 과제 점수는 전문가 평가 SJT($r$(48) = .39, $p$ = .006)와 보정된 상호평가 SJT 점수($r$(48) = .42, $p$ = .003)와 유의미한 정적 상관을 나타냈다.

세부 과제별로 살펴보면, 분석적 사고 과제 점수는 선다형 직무적성검사($r$(48) = .44, $p$ = .003)와 구성형 직무적성검사($r$(48) = .40, $p$ = .007) 점수와 유의미한 상관을 보였다. 의사소통능력 과제 점수는 전문가 평가 SJT($r$(48) = .46, $p$ = .001), 보정된 SJT 점수($r$(48) = .40, $p$ = .005)와 각각 유의미한 정적 상관을 나타냈다. 문제해결능력 과제 점수 또한 전문가 평가 SJT($r$(48) = .52, $p$ < .001), 상호평가 SJT($r$(48) = .46, $p$ = .002), 보정된 SJT($r$(48) = .56, $p$ < .001) 점수와 유의미한 정적 상관이 관찰되었다. 특히 분석적 사고 과제와 문제해결능력 과제 간 상관이 높게 나타나($r$(48) = .54, $p$ < .001), 두 과제 간의 밀접한 관련성을 시사하였다.

요약하면, 분석적 사고 과제는 직무적성검사 점수와 높은 상관계수를 보였으며, 의사소통능력 및 문제해결능력 과제는 SJT 점수와 높은 상관을 나타냈다. 이러한 상관분석은 이후 다변량 분석의 전제조건을 확인하기 위해 수행한 것이며, 과도한 상관은 발견되지 않아 후속 회귀분석 수행에 문제가 없음을 확인하였다.

표 1. 각 과제 점수 및 상관계수표

| | | *m* | *S.D.* | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X1 | 선다형 | 76.38 | 31.51 | | | | | | | | | |
| X2 | 구성형 | 87.02 | 27.59 | **.89** | | | | | | | | |
| X3 | 전문가평가 SJT | 8.46 | 1.42 | .21 | .24 | | | | | | | |
| X4 | 상호평가 SJT | 8.19 | 0.98 | -.01 | -.15 | **.59** | | | | | | |
| X5 | 보정된 SJT | 50.04 | 6.86 | .17 | .20 | **.79** | **.48** | | | | | |
| X6 | 업무 시뮬레이션 | 28.14 | 6.20 | .21 | .19 | **.39** | .24 | **.42** | | | | |
| X7 | 분석적사고 | 6.99 | 2.17 | **.44** | **.40** | .25 | .16 | .20 | **.57** | | | |
| X8 | 의사소통능력 | 15.13 | 4.07 | .23 | .18 | **.46** | .20 | **.40** | **.68** | .25 | | |
| X9 | 문제해결능력 | 7.49 | 2.02 | .21 | .18 | **.52** | **.46** | **.56** | **.75** | **.54** | .26 | |

Note. p < .05는 굵게 표시하였으며; 전문가평가 SJT와 상호평가 SJT 변수는 독창성(originality) 및 흐름(coherence) 기준 점수의 합으로 계산되었고; 의사소통능력 변수는 두 개의 세부 업무 시뮬레이션 과제 점수를 합산하여 산출되었다.

## 회귀분석

업무 시뮬레이션 과제에 대해 SJT와 직무적성검사를 예측변수로 선형 회귀분석을 실시한 결과는 다음과 같다. SJT 점수는 보정된 상호평가 점수를 사용하였으며, 직무적성검사 점수는 선다형과 구성형 점수를 활용하였다.

전체 업무 시뮬레이션 과제에 대한 회귀분석 결과, 모델은 유의미한 설명력을 나타냈다 ($R^2$ = .20, Adjusted $R^2$ = .14, $F(3, 43)$ = 3.58 ($p$ = .02). 상수는 유의하지 않았으며 ($b$ = 8.32, 95% CI [-4.63, 21.27], $t$ = 1.30, $p$ = .20), SJT 변인은 예측력이 유의미하였다 ($b$ = 0.36, 95% CI [0.11, 0.62], $\beta$ = 0.40, $t$ = 2.88, $p$ = .01). 이에 반해, 직무적성검사 구성형 변인과 선다형 변인은 예측력이 모두 유의

표 2. 업무 시뮬레이션 과제에 대한 회귀분석 결과

| predictor | | *b* | *b* 95% CI [LL,UL] | *B* | *t* | *p* |
|---|---|---|---|---|---|---|
| (Intercept) | | 8.32 | [-4.63 21.27] | | 1.30 | .20 |
| SJT | | 0.36 | [0.11 0.62] | 0.40 | 2.88 | .01 |
| 직무적성검사 | 구성형 | -0.02 | [-0.16 0.12] | -0.01 | -0.27 | .79 |
| | 선다형 | 0.04 | [-0.01 0.16] | 0.22 | 0.74 | .47 |
| $R^2$= 0.20, Adjusted $R^2$=0.14, Durbin-Watson=1.77, $F(3,43)$=3.58($p$ = .02) | | | | | | |

미하지 않았다. 구성형 변인의 회귀계수는 $b$ = -0.02 (95% CI [-0.16, 0.12], $\beta$ = -0.01, $t$ = -0.27, $p$ = .79)이었으며, 선다형 변인의 회귀계수는 $b$ = 0.04 (95% CI [-0.01, 0.16], $\beta$ = 0.22, $t$ = 0.74, $p$ = .47)이었다.

다중 공선성 분석 결과, 구성형 변인과 선다형 변인의 VIF 값은 각각 4.86과 4.80으로, 공선성 문제의 기준치인 5 미만을 만족하였다. 또한, Durbin-Watson 통계량 = 1.77로, 잔차 간 독립성이 확보된 것으로 나타났다.

이러한 결과는 SJT가 전반적으로 업무 시뮬레이션 과제를 더 잘 예측하는 반면, 직무적성검사 구성형 및 선다형 변인은 그렇지 않음을 보여 가설1을 지지한다.

업무 시뮬레이션의 세부 과제 중 하나인 분석적 사고 과제를 종속 변수로 설정하고, SJT와 직무적성검사를 예측 변수로 설정하여 선형 회귀분석을 실시한 결과는 다음과 같다.

분석적 사고에 대한 회귀분석 결과, 모델은 유의미한 설명력을 나타냈다 ($R^2$ = .21, Adjusted $R^2$ = .15, $F$(3, 43) = 3.59, $p$ = .02). 상수는 유의하지 않았다 ($b$ = 2.88, 95% CI [-1.35, 7.10], $t$ = 1.38, $p$ = .18).

SJT 변인은 유의미한 예측 변인이 아니었으며 ($b$ = 0.04, 95% CI [-0.04, 0.12], $\beta$ = 0.14, $t$ = 0.98, $p$ = .33), 구성형 변인과 선다형 변

표 3. 분석적 사고에 대한 회귀분석 결과

| predictor | | $b$ | $b$ 95% CI [LL,UL] | $B$ | $t$ | $p$ |
|---|---|---|---|---|---|---|
| (Intercept) | | 2.88 | [-1.35 7.10] | | 1.38 | .18 |
| SJT | | 0.04 | [-0.04 0.12] | 0.14 | 0.98 | .33 |
| 직무적성검사 | 구성형 | 0.00 | [-0.04 0.05] | 0.03 | 0.10 | .92 |
| | 선다형 | 0.03 | [-0.01 0.06] | 0.39 | 1.30 | .20 |
| $R^2$= 0.21, Adjusted $R^2$=0.15, Durbin-Watson=2.41, $F$(3,43)=3.59($p$= .02) | | | | | | |

표 4. 의사소통능력에 대한 회귀분석 결과

| predictor | | $b$ | $b$ 95% CI [LL,UL] | $B$ | $t$ | $p$ |
|---|---|---|---|---|---|---|
| (Intercept) | | 2.33 | [-6.17 10.83] | | 0.55 | .58 |
| SJT | | 0.23 | [0.06 0.40] | 0.39 | 2.78 | .01 |
| 직무적성검사 | 구성형 | -0.04 | [-0.12 0.06] | -0.24 | -0.79 | .44 |
| | 선다형 | 0.05 | [-0.03 0.13] | 0.38 | 1.28 | .21 |
| $R^2$= 0.20, Adjusted $R^2$=0.15, Durbin-Watson=1.52, $F$(3,43)=3.64($p$ = .02) | | | | | | |

인도 유의미하지 않았다. 구성형 변인의 회귀계수는 $b$ = 0.00 (95% CI [-0.04, 0.05], $\beta$ = 0.03, $t$ = 0.10, $p$ = .92)로 나타났고, 선다형 변인의 회귀계수는 $b$ = 0.03 (95% CI [-0.01, 0.06], $\beta$ = 0.39, $t$ = 1.30, $p$ = .20)로 나타났다.

다중 공선성 분석 결과, 구성형 변인과 선다형 변인의 VIF 값은 각각 4.66과 4.60으로, 공선성 문제의 기준치인 5 미만을 만족하였다. 또한, Durbin-Watson 통계량 = 2.41으로, 잔차 간 약한 음의 자기상관이 있으나 어느 정도 독립성이 확보된 것으로 나타났다. SJT와 직무적성검사 변인 모두 분석적 사고를 예측하지 못하기에, 가설1-a는 지지되지 않았다.

다음으로는 SJT와 직무적성검사를 예측변수로 의사소통능력 과제에 대해 회귀분석을 실시하였고, 그 결과는 다음과 같다.

의사소통능력 과제에 대한 회귀분석 결과, 모델은 유의미한 설명력을 나타냈다 ($R^2$= .20, Adjusted $R^2$= .15, $F$(3, 43) = 3.64, $p$ = .02). 상수는 유의하지 않았다 ($b$ = 2.33, 95% CI [-6.17, 10.83], $t$ = 0.55, $p$ = .58).

SJT 변인은 유의미한 예측 변인으로 나타났으며 ($b$ = 0.23, 95% CI [0.06, 0.40], $\beta$ = 0.39, $t$ = 2.78, $p$ = .01), 직무적성검사 구성형 변인과 선다형 변인은 유의미하지 않았다. 구성형 변인의 회귀계수는 $b$ = -0.04 (95% CI [-0.12, 0.06], $\beta$ = -0.24, $t$ = -0.79, $p$ = .44)이었고, 선다형 변인의 회귀계수는 $b$ = 0.05 (95% CI [-0.03, 0.13], $\beta$ = 0.38, $t$ = 1.28, $p$ = .21)로 나타났다.

다중 공선성 분석 결과, 구성형 변인과 선다형 변인의 VIF 값은 각각 4.86과 4.80으로, 공선성 문제의 기준치인 5 미만을 만족하였다. 또한, Durbin-Watson 통계량 = 1.52로, 잔차 간 독립성이 확보된 것으로 나타났다.

이러한 결과는 SJT가 의사소통능력 과제를 더 잘 예측하는 반면, 구성형 및 선다형 변인은 그렇지 않았음을 보여준다. 이는 가설1-b를 지지하는 결과다.

마지막으로, SJT와 직무적성검사를 예측변수로 문제해결능력 과제에 대해 회귀분석을 실시한 결과는 다음과 같다.

문제해결능력에 대한 회귀분석 결과, 모델은 유의미한 설명력을 나타냈다 ($R^2$ = .33, Adjusted $R^2$ = .28, $F$(3, 39) = 6.38, $p$ = .00). 상수는 유의하지 않았으나 ($b$ = -1.74, 95% CI [-6.23, 2.76], $t$ = -0.78, $p$ = .44), SJT 변인은

표 5. 문제해결능력에 대한 회귀분석 결과

| predictor | | $b$ | $b$ 95% CI [LL,UL] | $B$ | $t$ | $p$ |
|---|---|---|---|---|---|---|
| (Intercept) | | -1.74 | [-6.23 2.76] | | -0.78 | .44 |
| SJT | | 0.19 | [0.10 0.28] | 0.56 | 4.07 | <.001 |
| 직무적성검사 | 구성형 | -0.02 | [-0.06 0.03] | -0.19 | -0.70 | .49 |
| | 선다형 | 0.02 | [-0.02 0.06] | 0.24 | 0.91 | .37 |
| $R^2$= 0.33, Adjusted $R^2$=0.28, Durbin-Watson=2.35, $F$(3,39)=6.38 ($p$ = .00) | | | | | | |

유의미한 예측 변인으로 나타났다 (b= 0.19, 95% CI [0.10, 0.28], $\beta$ = 0.56, $t$ = 4.07, $p$ < .001). 이에 반해, 구성형 변인과 선다형 변인은 유의미하지 않았다. 구성형 변인의 회귀계수는 $b$ = -0.02 (95% CI [-0.06, 0.03], $\beta$ = -0.19, $t$ = -0.70, $p$ = .49)이었으며, 선다형 변인의 회귀계수는 $b$ = 0.02 (95% CI [-0.02, 0.06], $\beta$ = 0.24, $t$ = 0.91, $p$ = .37)로 나타났다.

다중 공선성 분석 결과, 구성형 변인과 선다형 변인의 VIF 값은 각각 4.11과 4.00으로, 공선성 문제의 기준치인 5 미만을 만족하였다. 또한, Durbin-Watson 통계량 = 2.35으로, 잔차 간 약한 음의 자기상관이 있으나 어느 정도 독립성이 확보된 것으로 나타났다.

SJT은 유의미한 예측 변인으로 문제해결능력을 더 잘 예측하는 반면, 구성형 및 선다형 변인은 그렇지 않음을 알 수 있고, 이는 가설 1-c를 지지한다.

### SJT 상호평가 점수의 타당도 분석

본 연구에서는 상호평가 기반 논술형 SJT의 준거타당도를 살펴보기 위해, 상호평가 채점 결과와 전문가 평정 결과 간의 상관을 확인하였다. 상호평가 원점수와 전문가 평정 점수 간 상관은 $r$ = .59 ($p$ < .001)이었으며, 보정 후 상호평가 점수와 전문가 평정 점수 간 상관은 $r$ = .79 ($p$ < .001)으로 나타났다. 이는 상호평가 방식이 전문가 평정과 높은 일치도를 보여, 상호평가 채점 결과가 준거 타당도가 높다는 근거가 되며 두 번째 가설을 지지하는 결과이다.

## 논 의

본 연구는 직무적성검사와 대비했을 때, 논술형 SJT가 업무 시뮬레이션 과제를 더 잘 예측하는지, 그리고 SJT의 상호평가 결과가 전문가 평정과 높은 일치도를 보이는 지를 검증하기 위해 수행되었다. 연구 결과, 논술형 SJT는 의사소통능력과 문제해결능력 과제에서는 물론 전반적 업무 시뮬레이션 과제에서도 직무적성검사보다 우수한 예측력을 보였다. 다만 분석적 사고력에 대해서는 유의미하게 예측하지 못했다. 또한, 상호평가 결과와 전문가 평정 결과 간 높은 상관도 관찰되었다.

이상의 결과가 조직 선발 장면에 주는 시사점은 두 가지다. 첫째, 논술형 SJT가 단순 지식이나 추론 능력 검증을 넘어, 실제 맥락에서 요구되는 의사소통능력, 혹은 문제해결능력 등 지원자의 역량을 포착하는 데 효과적이라는 것이다. 둘째, 상호평가는 큰 비용을 들이지 않고도 평가의 품질을 확보할 수 있는 현실적인 대안이 될 수 있다는 것이다. 이러한 결과는 조직에서 지원자의 실제 업무 수행을 보다 현실적으로 평가할 수 있는 선발 도구를 고민할 때, 기존 직무적성검사에만 의존하기보다는 논술형 SJT를 활용해 실질적인 역량 중심 평가로의 전환을 적극적으로 고려해야 함을 시사한다. 특히, SJT에 상호평가 방식을 접목할 경우 평가자의 부담과 비용을 완화하면서도 평가 신뢰도를 유지할 수 있어, 대규모 인원 선발 시 인사 실무자들이 직면하는 문제를 효과적으로 해결할 수 있다. 나아가 고객 응대가 핵심인 영업 · 서비스 직종과 같이 섬세한 의사소통 능력이 직무 성공에 직결되는 분야에서는 논술형 SJT를 중심 평가 도구로 활용하는 방안을 적극 고려

할 필요가 있다.

SJT는 직무적성검사가 예측하지 못한 의사소통능력을 유의미하게 예측하였다. 본 연구에서 사용된 의사소통능력 과제는 실무와 유사한 상황을 설정해 지원자가 맞닥뜨릴 수 있는 도전적인 의사소통 상황에 대한 대응을 요구하였다. 예를 들어, 인턴 사원에게 불합격 통보 메일을 작성하거나 부정 청탁을 거절하는 메일을 작성하도록 하여, 민감한 상황에서 적절하게 본인의 의사를 전달하는 능력을 평가하였다. 이와 같은 과제는 지원자가 실제 상황이었다면 심리적 불편함을 겪을 만한 상황에서 설득력을 발휘하도록 설계되었으며, 정서적 지능과 대인관계능력까지 반영하고자 하였다. 정서적 지능은 감정을 정확히 인식, 표현, 조절하는 능력으로, 논술형 SJT가 이러한 상황에서 유의미한 예측력을 보인다는 점은 단순한 의사소통능력을 넘어 복잡한 직무 상황에서 요구되는 대인 간 의사소통능력까지도 포함한다는 점에서 실제 직무수행능력을 평가할 수 있음을 의미한다. 따라서 SJT는 전통적인 직무적성검사에서 다루지 못하는 언어적 표현뿐만 아니라, 대인 간 의사소통능력을 평가하는 중요한 도구로 활용될 수 있다. 이는 고객과의 접점에서 섬세한 의사소통이 필수적인 영업직이나 고객관리 직군에서 SJT의 높은 실무적 활용 가능성을 보여준다.

SJT는 또한 직무적성검사가 예측하지 못한 문제해결능력도 유의미하게 예측하였다. 본 연구에서 사용된 문제해결능력 과제는 신입직원들의 높은 퇴사율 문제를 해결하기 위해 신입직원 만족도 문항을 작성하고, 우수 문항을 선정하는 것이었다. 이는 문제 상황을 분석하고 해결 방안을 탐색하는 과정을 필요로 하는데, SJT가 문제를 짚어내고 해결책을 제시하도록 요구함으로써 유사한 사고 과정을 요했을 것으로 예상할 수 있다. 반면, 직무적성검사는 짧은 응답을 요구하거나 작성하게 하는 방식이었는데, 이 과정에서는 지원자의 심층적 사고 과정을 파악하기 어려웠기 때문일 수 있다. 따라서 우선순위 설정이나 대안 비교가 요구되는 복합적 직무 상황에서는 SJT가 직무적성검사보다, 지원자의 역량을 더 잘 평가하는 것으로 판단된다.

반면, 분석적 사고력에서는 SJT와 직무적성검사 모두 예측력이 제한적이었다. 이런 결과가 관찰된 이유는, 두 검사 도구가 분석적 사고 과제에서 요구되는 정보 탐색, 분류, 매칭 등 종합적 판단 과정을 직접적으로 평가하기에 적합하지 않았기 때문일 수 있다. 분석적 사고 과제는 평판 정보 검토 후 조직 인재상과 매칭하며 불필요 정보를 제거하고 핵심을 종합・판단하도록 설계되었다. 그런데, SJT는 제한된 시나리오 내에서 적절한 대응을 서술하도록 구성되어 있어 외부 정보의 분류 및 매칭과 같은 심층적 절차를 직접 평가하기는 어려웠을 수 있다. 직무적성검사 역시 언어 및 수리 능력 평가에 초점을 두어, 복합적, 맥락적 정보 통합 과정을 필수적으로 요구하지 않는다. 따라서 두 검사 모두 분석적 사고 과제의 특수성을 충분히 반영하지 못해 예측 타당성에 한계가 있었을 수 있다는 것이다. 따라서 후속 연구에서는 분석적 사고를 정밀하게 평가할 수 있는 새로운 과제를 개발・검증할 필요가 있다.

또한, 본 연구에서는 상호평가 방식을 도입함으로써 전문가 채점자의 부담을 경감하면서도 높은 평가 타당도를 유지할 수 있다는 점을 확인하였다. 이는 특히 대규모 선발 상황에 실무적 시사점을 제공한다. 전통적으로 전

문가 평정은 높은 신뢰도와 타당도를 담보하지만, 많은 인원에게 논술형 답안이나 SJT 서술형 답변을 일일이 평가하게 하려면 시간과 비용이 상당히 요구된다. 상호평가 방식은 참여자가 평가를 분담함으로써 전문가에 대한 의존도를 낮추고, 조직 또는 교육 현장에서 현실적으로 수용가능한 평가가 이루어지게 할 수 있다. 나아가 평가자의 능력을 반영한 보정 절차를 도입한 결과, 상호평가 점수와 전문가 평가 간의 일치도가 높아짐을 확인하였다. 이러한 보정 절차는 실제 선발 상황에서도 적용 가능하며, 평가의 타당도를 강화하는 실질적 대안이 될 수 있다.

4차 산업혁명 시대에 부합하는 인재를 선발하기 위해서는 단순히 지식을 묻는 시험에서 벗어나, 직무와 관련된 문제해결능력, 비판적 사고와 같은 사고력을 측정할 수 있는 도구를 사용해야 한다. 기존 선다형 직무적성검사는 정답 선택에 집중하는 구조로 과거 지식 중심의 평가에는 적합했지만, 오늘날의 정보 활용 능력을 평가하기에는 한계가 있다.

이런 맥락에서 본 연구가 선발 방식의 변화에 긍정적 영향을 줄 수 있을 것으로 예상되지만, 몇 가지 한계점도 있다. 첫째, 본 연구에서 사용된 업무 시뮬레이션 과제는 인사팀 직원 역할에 초점을 맞춘 시나리오로 구성되었으므로 모든 직무에 동일하게 적용하기에는 제약이 있을 수 있다. 다만 선행연구와 직무분석 결과를 토대로 분석적 사고, 의사소통능력, 문제해결능력이라는 핵심 역량을 묻는 세부 과제들을 설계했기 때문에 일반화가 어렵지는 않으나, 다양한 직무 및 산업 맥락에서 추가 검증이 필요하다. 둘째, 본 연구는 피로도 효과를 고려하여 두 시간 안에 세 과제를 수행하도록 설계했으나 실제 선발 과정에서는 이보다 긴 시간이나 단계별 평가가 이루어질 수 있다. 따라서 향후 연구에서는 과제 순서나 자극 유형을 다양화하고 피로도 관리 방안을 병행하여 실무 환경에 보다 근접한 평가 조건을 마련할 필요가 있다. 마지막으로, 분석적 사고 과제에서 예측력이 제한적이었으므로 심층 정보 탐색과 분류, 매칭 등 종합적 판단 과정을 측정할 수 있는 평가 도구를 개발하고 검증해야 할 것이다.

이상의 한계에도 불구하고 본 연구는 타당도 관점에서 논술형 SJT의 효과를 실증적으로 탐구했다는 점에서 의의를 찾을 수 있다. 논술형 SJT가 기존 지식 중심 검사만으로는 포착하기 어려운 의사소통능력과 문제해결능력을 유의미하게 예측한다는 결과와 상호평가 방식의 도입 가능성을 확인한 결과는 고무적이다. 이와 관련된 후속 연구가 더 활발해질 것을 기대해본다.

## 참고문헌

권은아, & 이종구 (2021). 정직성 상황판단검사가 과업수행 및 반생산적 업무수행에 미치는 효과: 도덕적 해이의 매개효과. 한국심리학회지: 산업 및 조직, 34(3), 541-561.

김의수, 김명소 & 한영석 (2011). 상황판단검사의 채점용 답 결정방식과 채점방식이 타당도에 미치는 영향: 왜곡여부에 따른 비교 연구. 한국심리학회지: 산업 및 조직, 24(1), 231-255.

박정애, & 박주용 (2019). 동료평가 정확도 향상 방안의 비교: 평가 기준에 대한 학생들 간 토론 대 전문가 평가 사례 제시.

*Korean Journal of Cognitive Science, 30*(4).

오인수, & 서용원 (2002). 구조화된 역량기반 채용면접 체계: 이론적 고찰과 개발 사례. 인적자원 개발연구, 4(1), 49-75.

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

이승철, & 어수봉 (2017). 신입직원 선발시스템 개선에 관한 탐색적 연구: H공사 사례를 중심으로. 경영컨설팅연구, 17(1), 255-268.

정현선, & 박동건 (2009). 지식형 상황판단검사의 채점용 답(Scoring key) 결정과 채점 방식에 따른 준거관련 타당도 비교 연구. 한국심리학회지: 산업 및 조직, 22(1), 53-70.

조성준, & 박찬균 (2020). 채용도구로서 NCS의 타당성 검토: 직업기초능력 영역을 중심으로. *HRD 연구, 22*(2), 43-75.

한국산업인력공단 (2020). *NCS* 기반 직업기초능력 모형 재설계 연구. 한국산업인력공단.

미국 교육부 (n.d.). *Employability skills framework*. https://cte.ed.gov/initiatives/employability-skills-framework

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*(1), 125-154. https://doi.org/10.1111/j.1744-6570.2003.tb00146.x

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology*, *78*(3), 387-409. https://doi.org/10.1348/096317905X26994

Bock, L. (2015). *Work rules! Insights from inside Google that will transform how you live and lead*. Hachette Books.

Brannick, M. T., & Ellingson, J. E. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*(2), 104-116.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. *Cambridge University Press.*

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence, 43*(1), 52-64. https://doi.org/10.1016/j.intell.2013.08.004

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3), 287-322. https://doi.org/10.3102/00346543070003287

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, 72(3), 493-511. https://doi.org/10.1037/0021-9010.72.3.493

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24*(1), 79-132. https://doi.org/10.1016/S0160-2896(97)90014-3

Hermelin, E., Lievens, F., & Robertson, I. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment 15*(4), 405-411.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*(1), 72-98. https://doi.org/10.1037/0033-2909.96.1.72

Iqbal, M. Z., Ivan, R., Robb, C., & Derby, J. (2025). Evaluating factors that impact scoring an open response situational judgment test: a mixed methods approach. *Frontiers in Medicine, 11,* Article 1525156. https://doi.org/10.3389/fmed.2024.1525156

Kasten, N., Freund, P. A., & Staufenbiel, T. (2020). "Sweet little lies": An in-depth analysis of faking behavior on Situational Judgment Tests compared to personality questionnaires. *European Journal of Psychological Assessment, 36*(1), 136-148. https://doi.org/10.1027/1015-5759/a000479

Krause, D. E., & Gebert, D. (2003). A comparison of assessment center and checklist ratings in a leadership simulation. *International Journal of Selection and Assessment, 11*(2 – 3), 177-183. https://doi.org/10.1111/1468-2389.00240

Landers, R. N., & Sanchez, D. R. (2022). Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation. I*nternational Journal of Selection and Assessment, 30*(1), 1-13. https://doi.org/10.1111/ijsa.12376

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2016). Does peer assessment promote student learning? *A meta-analysis. Assessment & Evaluation in Higher Education, 41*(6), 869-900. https://doi.org/10.1080/02602938.2015.1049384

Liu, N.-F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279 – 290.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63-91. https://doi.org/10.1111/j.1744-6570.2007.00065.x

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. J*ournal of Applied Psychology, 86*(4), 730-741. https://doi.org/10.1037/0021-9010.86.4.730

McDaniel MA, Nguyen NT. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*, 103-113.

Mortaz Hejri, S., Ho, J. L., Pan, X., Park, Y. S., Sam, A. H., Mangardich, H., et al. (2023). Validity of constructed-response situational judgment tests in training programs for the health professions: A systematic review and meta-analysis protocol. *PLoS ONE, 18*(1), e0280493. https://doi.org/10.1371/journal.pone.0280493

NNBIA. (2014). *Common employability skills: A*

*foundation for success in the workplace*. http://www.nam.org/Issues/Workforce-Development/Common-Employability-Skills

NESTA. (2017). *The future of skills: Employment in 2030*. https://www.nesta.org.uk/report/the-future-of-skills-employment-in-2030/

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2015). Situational judgment tests in medical education: Research, theory and practice. *Medical Teacher, 37*(6), 524-538. https://doi.org/10.3109/0142159X.2014.1003054

Patterson, F., Zibarras, L. D., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Medical Teacher, 38*(1), 3-17. https://doi.org/10.3109/0142159X.2015.1072619

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & De Fruyt, F. (2003). A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology, 88*(6), 1068-1081. https://doi.org/10.1037/0021-9010.88.6.1068

Salgado, J. F., & Moscoso, S. (2019). Meta-analysis of the validity of general mental ability for supervisory performance. *International Journal of Selection and Assessment, 27*(3), 254-263.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin, 124(2), 262-274. https://doi.org/10.1037/0033-2909.124.2.262

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on workforce productivity. Journal of Applied Psychology, 64(6), 609-626. https://doi.org/10.1037/0021-9010.64.6.609

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed., pp. 553-581). New York, NY: The Guilford Press.

Schunn, C., Godley, A., & DeMartino, S. (2016). The reliability and validity of peer review of writing in high school AP English classes. *Journal of Adolescent & Adult Literacy, 60*(1), 13-23.

Shvetsova, O. (2025). Managing Global Talent: Innovative Solutions and a Sustainable Strategy Using a Human-Centric Approach (Case of Republic of Korea). *Administrative Sciences, 15*(5), 190. https://doi.org/10.3390/admsci15050190

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Tam, P., & Delaney, K. (2005). Google's Quest to Build a Better Boss. The Wall Street Journal.

U.S. Department of Labor Employment and Training Administration. (1981). General Aptitude Test Battery: Manual and research report. *U.S. Department of Labor*.

Van Iddekinge, C. H., Lievens, F., & Sackett, P.

R. (2023). Personnel selection: A review of ways to maximize validity, diversity, and the applicant experience. *Personnel Psychology, 76*(2), 651-686. DOI: 10.1111/peps.12578.

투고일자: 2025-04-07
수정일자: 2025-06-24
게재일자: 2025-08-03

한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2025. Vol. 38, No. 3, 313-334
https://doi.org/10.24230/kjiop.v38i3.313-334

# Predictive Validity of Peer-Assessed Constructed SJT and Aptitude Tests on Work Simulation Performance

**Songeui Kim** **Jooyong Park**

Dapartment of Psychology, Seoul National University

The use of valid selection tools is critical for identifying high-potential talent and ensuring effective organizational hiring decisions. This study empirically investigates the predictive validity of a constructed situational judgment test (SJT) relative to a general aptitude test using work simulation tasks designed to reflect actual job demands. Four simulation tasks were specifically developed to assess analytical thinking, communication skills, and problem-solving abilities. Fifty university students completed both the constructed SJT and a standardized general aptitude test. Regression analyses were conducted using their performance on the simulation tasks as dependent variables. Results indicated that the writing-based SJT significantly predicted performance in communication and problem-solving simulation tasks, whereas the general aptitude test showed insignificant predictive validity across these domains. Additionally, peer-assessed SJT scores correlated strongly with expert ratings, further supporting the criterion validity of peer evaluations within the SJT context. These findings underscore the promise of writing-based assessments, particularly those utilizing peer evaluation, as effective tools for assessing cognitive skills in selection settings. This study contributes to the literature advocating for cognitively demanding assessment formats and offers practical insights into developing cost-effective, scalable hiring systems.

*Key words : SJT, peer assessment, general aptitude test, work simulation, validity, talent selection*