Charting the Future of Assessments

Kyllonen, P., Sevak, A., Ober, T., Choi, I., Sparks, J., & Fishtein, D.

12/31/24

Table of contents

평가의 미래방향 설정

Authors

Patrick Kyllonen, Amit Sevak, Teresa Ober, Ikkyu Choi, Jesse Sparks, & Daniel Fishtein
(ETS Research Institute, ETS, Princeton, New Jersey United States)

평가는 개인(또는 집단)의 기술, 행동, 성향 또는 기타 속성을 측정하거나 평가하기 위한 다양한 접근 방식을 포함하는 포괄적인 개념이다. 평가는 입학시험, 직원 선발, 자격시험과 같은 표준화된 시험부터 국내 및 국제 수준에서 인지 및 행동적 기술을 평가하는 대규모 평가, 그리고 K-12 교육과정에서 활용되는 형성 평가에 이르기까지 다양하게 이루어진다.

이러한 다양한 유형의 평가는 광범위한 목적을 위해 사용되지만, 신뢰도, 타당도, 공정성과 같은 표준을 갖춘다는 공통점을 가진다. 심지어 교실에서 이루어지는 평가도 이러한 기준을 따른다.

우리는 미래의 평가는 어떤 기술을 측정할 것인가에 대한 강조점 변화, 측정 방식의 혁신, 시험 운영에 있어 첨단 기술의 활용, 그리고 평가를 통해 피험자가 받을 수 있는 정보의 가치와 종류의 확대를 포함하게 될 것이라고 본다.

이 논문에서 우리는 미래의 평가가 도전 과제를 포함하지만, 동시에 유망한 가능성을 지닌다고 주장하며 이에 대한 근거를 제시한다. 주요 도전 과제로는 보안 및 개인정보 노출 위험, 시험 점수 편향, 부적절한 시험 활용 등이 있으며, 이러한 문제들은 인공지능(AI)의 확산으로 인해 더욱 심화될 수 있다. 반면, 평가는 개인이 교육 및 경력 목표를 달성하고, 전반적인 삶의 질과 웰빙에 기여할 수 있는 기회를 확대할 가능성도 가지고 있다.

이러한 가능성을 실현하기 위해, 우리는 교육 및 직업 학습에서의 측정 과학에 기반한 증거 중심 접근 방식에 초점을 맞추고자 하며, 이는 본 논문의 핵심 주제이다.

주요어: 평가, 미래, 노동력, 기술 향상, 진로 선택, 인지 기술, 행동 기술, 소프트 스킬, 지속 가능한 기술, 표준화된 시험, 입학시험, 직원 선발, 자격시험, 국내 대규모 평가, 국제 대규모 평가, 표준, 신뢰도, 타당도, 공정성, 개별 지도, 피드백

doi:10.1002/ets2.12388

1 요약

우리는 평가의 미래가 다음과 같은 변화를 수반할 것으로 생각합니다: 측정하게 될 능력에 대한 강조점의 변화, 이러한 능력을 측정하는 방식의 혁신, 시험 운영을 위한 첨단 기술의 활용, 그리고 평가 응시자들이 받게 될 정보의 가치와 종류의 확장입니다. 평가는 수세기 동안 우리와 함께 해왔으며, 앞으로도 계속될 것입니다. 왜냐하면 시험과 평가는 의사결정을 지원하는 데 있어 효율적이고 증거 기반적인 방식으로 가치를 제공하기 때문입니다. 평가는 응시자, 학부모, 교사, 교육 행정가, 고용주, 연구자, 정책입안자 등 다양한 이해관계자들에게 능력에 대한 유용한 정보를 제공합니다. 평가는 특히 그들의 성취나 잠재력이 인정받지 못했을 수 있는 사람들에게 기회를 제공합니다. 시험은 응시자들에게 자신의 현재 위치와 향상을 위해 무엇을 해야 하는지에 대한 피드백을 제공함으로써 더 큰 가치를 창출합니다.

1.1 미래를 위한 능력: 기술 발전의 영향

지난 세기 동안 평가 분야의 노력과 발전은 주로 교육과정 능력—수학, 읽기, 과학 등 전통적인 K-12 교육과정이 목표로 하는 능력—의 평가와 관련되어 왔습니다. 이러한 능력들은 계속해서 중요할 것이지만, 다른 종류의 능력—협력, 문제해결, 비판적 사고, 창의성, 호기심, 직업윤리—의 중요성에 대한 인식이 높아지고 있습니다. 이들은 지속가능한 능력으로, 모든 종류의 교육, 훈련, 직무 및 맥락에서의 일반화 가능성과 유용성을 나타내며, 동시에 평가의 어려움을 반영하는 측정하기 어려운 능력입니다. 기술과 AI의 발전은 어떤 능력이 가장 가치 있는지를 변화시킬 것입니다. 대학원생 이상 수준의 언어, 예술 창작, 코딩에 대한 AI의 능력은 평가에 있어 도전과 기회를 제시합니다. 도전 과제는 자기평가 등급이 앞으로 더욱 중요해질 능력에 대한 유용한 정보를 제공하는 과제에 충분하지 않다는 것입니다. 기회는 수학, 읽기, 과학을 오늘날 측정할 수 있는 것과 같은 수준의 정교함으로 측정하기 어려운 구인들을 평가할 수 있도록 새롭고 혁신적인 평가 방법을 개발할 수 있다는 것입니다.

1.2 혁신적 측정: 측정하기 어려운 능력 평가를 위한 새로운 접근

측정하기 어려운 능력을 측정하는 주된 방법은 등급척도 방식입니다. 하지만 우리는 자기보고식 방법을 개선할 수 있습니다. 타인 보고는 준거 편향과 같은 자기보고 편향에 덜 취약하지만, 후광효과와 같은 자체적인 한계가 있습니다. 강제선택형 측정도 자기보고 편향을 줄입니다. 상황판단검사는 많은 측정하기 어려운 능력에 적용할 수 있는 유연한 측정

방법입니다. 평가의 미래는 자기보고에서 이러한 다른 형태의 측정으로 이동할 것 같습니다. 더 중요한 변화는 게임과 실제 협상 세션이나 협력적 문제해결 과제와 같은 상호작용적 과제와 같은 수행기반 측정의 개발과 채택을 포함할 것입니다. 성격의 수행측정은 오랫동안 추구되어 온 목표였으며, 수행측정은 등급평정에 비해 상당한 장점이 있습니다: 평정 편향에 취약하지 않으며 행동에 대한 주관적 평가가 아닌 객관적 표본이 될 수 있습니다. 그러나 많은 중요한 구인에 대해 수행측정이 아직 잘 개발되지 않았습니다. 우리는 수행측정이 과정 분석과 데이터 마이닝을 포함하는 무검사 측정으로 보완될 것이라고 믿으며, 이는 사용자나 학생 또는 직원의 능력 수준에 대한 추론을 도출하는 데 사용될 수 있습니다. 사회정서학습부터 학업 수행과 STEM 직무 참여에 이르기까지 다양한 영역에서 좋은 사례들이 있습니다.

1.3 운영 혁신: AI와 기술 기반 발전

검사 운영은 검사의 목적과 행정적 조건 및 제약사항, 문항 개발, 검사 구성, 보안, 품질 관리, 채점, 검사 평가를 포함하며, 이는 검사 산업의 핵심입니다. 검사를 타당하고, 신뢰할 수 있으며, 공정하고, 응시자와 다른 이해관계자들에게 유용하게 만드는 것과 관련된 운영에는 많은 도전적인 문제들이 있습니다. 검사의 시작부터 기술이 그래왔듯이, 대규모 언어 모델(LLM)과 다른 AI 기술을 포함한 기술의 발전이 검사 운영에 극적인 영향을 미칠 것 같습니다. 우리는 검사가 개발, 구성, 채점되는 방식과 관련된 효율성과 품질의 상당한 발전을 보게 될 것이며, 이는 안전하게 만들어지고 공정하게 되어 모든 응시자들이 검사의 가치를 볼 수 있고 검사 점수를 기반으로 한 추론이 적절하고 정당화된다고 확신할 수 있게 될 것입니다.

1.4 피드백: 학습과학 기반 통찰과 응시자를 위한 실행 계획

평가와 검사는 흔히 제공되는 검사 점수와 규준 및 기준점을 넘어서 응시자들에게 유용한 정보를 제공할 수 있습니다. 형성평가와 검사 연습 모두 학습에 상당하고 긍정적인 효과를 제공합니다. 인간 교수는 가장 강력한 교육적 개입 중 하나로 밝혀졌습니다. 컴퓨터 기반 교수도 마찬가지로 강력한 개입입니다. 교수는 피드백과 안내된 프롬프트를 제공하고 상호작용과 건설적인 행동을 장려합니다. 교수는 학습자에 대한 인지진단을 수행하며, 검사도 마찬가지입니다. 인지진단모델링은 AI 발전을 활용하고 교수의 개별화를 통해 학습자들에게 유용한 도움을 제공합니다. 피드백은 개별화를 통해 학습을 향상시키는 강력한 수단입니다; 생성형 AI는 학습자와 학생들에게 유용한 피드백을 제공할 수 있으며, 이는 유망한 새로운 방향입니다. 피드백, 교수, 학습자 지도 수립에 있어 증거 기반 학습 원리의 사용은 평가의 가치를 크게 향상시킬 것입니다. 적절하게 설계되고 개별화된 피드백의 제공은 교육의 형평성 목표를 달성하고 모든 학습자의 학습과 수행을 촉진할 수 있습니다.

1.5 요약 및 결론

기술과 AI의 발전은 측정할 능력, 측정 방법, 응시자와 이해관계자들에게 결과를 보고하는 방식, 그리고 결과 수령자들이 그 결과로 할 수 있는 것에 이르기까지 평가의 모든 측면에 깊은 영향을 미칠 것입니다. 소프트 스킬, 지속가능한 능력, 복합적 능력의 핵심 집합이 미래에 점점 더 중요해질 것 같습니다. 생애주기에 걸쳐 나타나는 능력의 이러한 증가하는 역할과 함께, 능력 개발을 평가하고 인정하는 시스템이 자리잡게 될 것입니다. 비학위 자격증은 능력을 보여주는 가치 있는 방법이 될 것입니다; 이러한 자격증은 대학에서 나올 수 있지만 기업이나 표준화된 시험 또는 학습평가 기관에서 나오더라도 동등하게 가치 있는 것으로 취급될 것입니다. 능력 습득의 인증을 얻기 위해 평가에 의존하는 것은 이러한 인증에 대한 보안 문제의 중요성을 높일 것입니다. 우리는 미래에 점점 더 중요해질 것 같은 많은 능력에 대해 좋은 평가를 설계해야 할 것입니다. 평가에 대한 태도는 매우 긍정적입니다: 평가는 시험 사용자들이 새로운 능력을 습득하도록 동기를 부여하고 기회를 추구하고 경력을 발전시킬 준비가 되었다고 자신감을 느끼게 해주며, 이는 AI가 주도하는 직장의 변화와 함께 점점 더 중요해질 것입니다.

1.6 권고사항

우리는 몇 가지 권고사항을 제시합니다. 첫째, 능력의 변화하는 특성을 모니터링해야 합니다—노동시장에서 요구되는 능력은 교육 표준과 교육과정에 영향을 미치므로, 이러한 변화를 예측하는 것이 유용합니다. 둘째, 우리는 협력적이고 다중양식적 접근을 포함하여 더 풍부한 평가 방법과 새롭고 혁신적인 접근을 계속 추구해야 합니다. 셋째, 문항 개발, 개별화, 채점, 보안, 결과 보고와 같은 검사 운영의 다양한 측면은 기술과 AI의 급속한 발전에 영향을 받고 있으며, 그 속도가 늦춰질 것 같지 않으므로 우리는 이러한 변화에 신속히 대응해야 합니다. 마지막으로, 우리는 응시자들에게 그들이 어디에 있고 어떻게 향상될 수 있는지에 대한 통찰을 제공하기 위해 유용하고 실행 가능한 피드백을 계속 제공해야 합니다.

ETS 연구소는 네 가지 연구 분야를 통해 이러한 방향에 대응하고 있습니다. 이는 평가의 개별화; 혁신적이고 상호작용적인 디지털 평가 제작을 위한 설계 원칙 개발; 자동화된 콘텐츠 생성과 채점을 포함한 책임있고 윤리적인 AI 응용을 위한 표준 개발; 그리고 격차를 해소하는 차세대 교육 시스템의 개념화를 통한 정책과 실천에의 영향에 초점을 맞추고 있습니다. 여기서 개괄된 연구와 ETS 연구소의 연구 분야를 통해, 우리는 성취와 개발된 능력을 측정하는 전통적인 역할을 포기하지 않으면서도 인간의 학습을 더 잘 지원하도록 평가를 재목적화할 수 있는 위치에 있습니다.

이 비전의 달성을 가능하게 할 교육과 능력 평가의 발전을 촉진하기 위해, 우리는 상당한 연구 투자를 요청합니다. 전 세계 교육 지출은 연간 5조 달러 이상으로, 전 세계 GDP의 약 6%입니다. 하지만 그 투자의 작은 부분만이 인간의 학습을 지원하고 교육적 진보를 모니터링하는 데 필요한 평가와 관련되어 있습니다. 초점과 투자를 통한 평가의 발전은 평가가 인간의 학습을 더 잘 지원한다는 비전을 달성하는 데 더 가까이 다가가는 데 중심적 역할을 할 것입니다.

2 평가의 미래 방향 설정

2.1 평가의 가치와 효용

이 논문의 첫 번째 부분에서는 평가(시험과 같은 것)의 중요성을 보여주는 증거를 살펴보고, 기술(skill)이 점점 더 중요한 가치가 되는 시대에 평가의 역할이 어떻게 변화할 수 있는지에 대해 이야기합니다. 평가의 다양한 목적—입시나 채용 시험 같은 중요한 평가부터 학습을 돕는 작은 시험까지—를 논의하고, 평가에 대한 인식 문제, 평가의 초점, 공정성과 신뢰성 등의 도전 과제도 살펴봅니다. 마지막으로, 평가가 미래에 어떻게 발전할 수 있을지 이야기하며, 학생들에게 유용한 정보를 제공하는 것, 중요한 기술을 찾고 측정하기 어려운 능력을 평가하는 방법을 발전시키는 것, 그리고 개별적인 피드백(맞춤형 조언)을 제공하는 것의 중요성을 강조합니다. 이후의 논문에서는 이러한 주제를 더 깊이 다룰 것입니다. 이 논문은 교육과 직업 분야에서 평가를 연구하는 학자들뿐만 아니라, 정책을 만드는 사람들과 재정을 지원하는 기관을 위한 것입니다. 우리는 전문적인 내용을 다루면서도 다양한 사람들이 이해할 수 있도록 쉽게 설명하려고 노력했습니다.

평가는 수세기 동안 우리와 함께 해왔으며 앞으로도 그럴 것입니다. 표준화 검사는 기원전 3세기까지 거슬러 올라가는데(Wainer, 1987), 당시 중국의 지원자들은 중국 황제의 보좌관이 되기 위해 음악, 궁술, 산술 및 기타 과목의 시험에 합격해야 했습니다(Himelfarb, 2019). 나폴레옹은 재능을 찾고 족벌주의를 피하기 위해 검사와 시험을 채택하여 다양한 분야에 걸쳐 공과대학으로 이어지는 에콜 폴리테크닉을 설립함으로써 고등교육에 혁명을 일으켰습니다(Bradley, 1975). ETS는 초대 회장 Henry Chauncey가 언급했듯이, "명문학교" 출신만이 아닌 "잘 알려지지 않은 고등학교의 자격 있는 사람들"을 찾기 위해 설립되었습니다(Lewin, 2002). 오늘날, 학교들은 계속해서 시험과 평가를 사용하지만, 다른 분야에서도 사용됩니다. 기업들은 채용, 리더십 개발, 기술 능력 인증에 평가를 사용합니다. 정부와 전문가 협회는 특히 경제의 핵심 부분에서 능력의 면허와 인증의 필요성을 인식하여 평가를 사용합니다. U.S. Congress, Office of Technology Assessment(1992)에서 설명된 것처럼, 역사를 통해, 그리고 전 세계적으로—중국, 러시아, 프랑스—평가는 다양한 목적으로 사용되어 왔으며, 이는 본 논문의 중요한 주제입니다.

검사와 평가가 지속된 이유는 의사결정을 지원하는 데 있어 효율적이고 증거 기반적인 방식으로 가치를 제공하기 때문입니다. 검사와 평가는 응시자, 학부모, 교사, 교육 행정가, 고용주, 연구자, 정책입안자와 같은 다양한 이해관계자들에게 수험자의 능력에 대한 유용한 정보를 제공합니다(Brookhart et al., 2020). 검사가 없는 세상은 대신 현재 평가 데이터에 의존하는 결정들을 위해 구시대적인 네트워크에 의존할 수 있습니다. 다른 방법들은 문제가 있습니다. 미국과 세계의 많은 지역에서 성적은 특히 비STEM 분야에서 점점 더 인플레이션되어 지원자들에 대한 정보를 덜 제공합니다(Ahn et al., 2019). 비학문적

자격증(예: 이력서; Kessler et al., 2019)은 조작 가능하고, 불공정하며, 특권층에게 유리합니다(Chetty et al., 2023). 면접은 성별, 인종/민족, 외모 편향이 개입될 수 있게 합니다(Chamorro-Premuzic, 2021). ChatGPT와 같은 생성형 AI 제품은 지원자의 지식, 기술, 능력, 경험의 지표로서 대학 에세이, 이력서 및 기타 서면 평가 형태의 타당도를 위협합니다. 전 세계적으로 이동성이 증가하고 극심한 인재 부족 현상이 있는 상황에서, 평가는 능력과 지식을 검증하는 효율적이고 경제적인 방법을 제공합니다—한 국가의 간호사가 다른 국가에서 자격 있는 역량의 증거를 제시할 수 있습니다. 평가의 역사를 통해 형평성과 효율성 속성 사이에 지속적인 긴장이 있어왔으며, 이는 본 논문 전체에서 다시 다루는 주제입니다.

평가는 특히 그들의 성취나 잠재력이 인정받지 못했을 수 있는 사람들에게 기회를 제공합니다(Schmill, 2022). 검사는 응시자들에게 그들의 현재 위치와 향상을 위해 다음에 무엇을 해야 하는지에 대한 피드백을 제공함으로써 더 큰 가치를 제공합니다(Wisniewski et al., 2020).

2.2 기술(역량): 미래의 새로운 화폐

경제협력개발기구(OECD)의 교육 및 기술 담당 이사이자 교육 정책 특별 고문인 안드레아스 슐라이허(Andreas Schleicher)는 "기술이 점점 화폐와 같은 역할을 하게 될 것"이라고 주장했습니다(ETS, 2023a).

ETS의 Human Progress Study 설문조사에서는 미래의 평가 방식에 대한 여러 질문이 포함되었습니다. 표 1에 따르면, 많은 응답자가 대학 학위보다 특정 기술을 증명하는 것이 더 중요해질 것이며, 마이크로크리덴셜(소규모 인증)이 이러한 기술을 증명하는 수단이 될 것이라고 동의했습니다. 또한, 이 조사에서 중소득 국가와 젊은 세대의 응답자들이 특히 이러한 변화에 강하게 공감하는 것으로 나타났습니다.

또한 표 2에서는 대학뿐만 아니라 기업 교육 및 시험 기관을 포함한 다양한 인증 기관이 발급하는 자격증이 대체로 비슷한 가치를 가지게 될 것이라는 응답이 많았습니다.

이처럼 기술과 그 인증에 대한 관심은 미래의 평가 방식에 영향을 미칠 중요한 요소와도 연결됩니다. 그것은 바로 평생 학습의 중요성이 점점 커지고 있다는 점입니다. OECD(2021)는 평생 학습을 "인생 전반에 걸쳐 이루어지는 모든 형태의 기술 개발과 지식 습득"이라고 정의했습니다. 표 3에 따르면, 응답자들은 지속적인 학습이 이제는 필수적이며, 단순히 경제적 안정뿐만 아니라 삶의 만족과 행복을 위해서도 필요하다고 생각하고 있습니다. 기업들도 직원들의 생산성을 높이기 위해 지속적인 직무 교육에 투자할 것이며, 이것 역시 평생 학습의 일부로 볼 수 있습니다.

Table 2.1: 자격 인증과 관련된 평가의 미래 예측

예측	동의 + 매우 동의	매우 동의
미래에는 특정 능력의 증명이 대학 학위보다 더 중요해질 것이다.	78%	32%
미래에는 마이크로 자격증(단기, 집중 인증)이 능력을 보여주는 가치 있는 방법이 될 것이다.	81%	27%

주: 데이터는 ETS 인간 진보 연구(ETS, 2023a)에서 가져왔음. 설문 문항: "다음 진술에 얼마나 동의하거나 동의하지 않습니까? (매우 동의하지 않음/다소 동의하지 않음/다소 동의/매우 동의)"

Table 2.2: 다양한 인증 출처의 가치

인증 출처	다소 또는 매우 가치있음
대학교	83%
기업 또는 기업 교육 프로그램	82%
산업별 인증 기관	82%
기술 기업	81%
공식 표준화 시험 또는 학습 평가 기관	80%
신뢰할 수 있는 온라인 학습 플랫폼	80%
산업 협회	79%
정부	77%
비영리 기관	71%

주: 데이터는 ETS 인간 진보 연구(ETS, 2023a)에서 가져왔음. 설문 문항: "다음 각각으로부터 자격증이나 인증을 받는 것이 얼마나 가치 있을 것 같습니까? (10점 척도: 1 = 전혀 가치 없음, 10 = 매우 가치 있음)"

Table 2.3: 지속적 학습의 중요성

진술	동의
지속적 학습은 삶을 더 충만하게 만든다.	87%
지속적 학습은 웰빙에 필수적이다.	86%
지속적 학습은 오늘날 세계에서 재정적 안정을 위해 필요하다.	86%
빠르게 변화하는 세상에서 지속적 학습은 이제 규범이다.	86%
지속적 학습은 과거 어느 때보다 지금 더 중요하다.	85%

주: 데이터는 ETS 인간 진보 연구(ETS, 2023a)에서 가져왔음. 설문 문항: "다음 진술에 얼마나 동의하거나 동의하지 않습니까? '지속적 학습'이란 전통적인 학교 교육 환경 이외에서 이루어지는 학습으로, 나중의 삶에서도 지속되는 것을 의미합니다. 여기에는 직업이나 여가를 위한 새로운 기술 학습, 특정 주제에 대한 지식이나 교육 확장 등이 포함될 수 있습니다. (매우 동의하지 않음/다소 동의하지 않음/다소 동의/매우 동의)"

Figure 2.1: 시험 응시 이유별 응답자 비율

주: 데이터는 ETS 인간 진보 연구(ETS, 2023a)에서 가져왔음. 설문 문항: '다음 중 학습 평가를 받고 싶은 이유는 무엇입니까? 해당하는 것을 모두 선택하세요.'

2.3 평가의 다양한 목적

평가는 여러 상황에서 다양한 이유로 사용됩니다. 그림 1은 ETS의 Human Progress Study(ETS, 2023a)에서 응답자들이 학교 입학이나 취업 선발과 같은 필수적인 이유 외에 시험을 치르는 다양한 이유를 선택한 비율을 보여줍니다. 이러한 이유는 지속적인 기술 향상, 현재 기술 수준 및 강점 파악, 새로운 분야에서의 잠재력 발견 등으로 다양합니다.

평가의 가치를 고려할 때, 그 의도된 사용 목적을 신중하게 생각하는 것이 중요합니다. 이 원칙은 교육 및 심리 평가 표준(AERA et al., 2014)에 명시되어 있으며, 여기서는 타당성을 "시험 점수의 해석이 제안된 사용 목적에 대해 증거와 이론으로 뒷받침되는 정도"로 정의하며, 타당성이 시험 개발 및 평가에서 가장 근본적인 고려 사항이라고 주장합니다(p. 11; 강조는 우리 것).

중요한 구분은 고부담(high-stakes) 평가와 저부담(low-stakes) 평가의 사용입니다. 이는 그림 2에 정의되어 있으며, National Research Council (1999a)에서 논의되었습니다.

High-stakes test: A test used to provide results that have important, direct consequences for individuals, programs, or institutions involved in the testing.
Low-stakes test: A test used to provide results that have only minor or indirect consequences for individual, programs, or institutions involved in the testing.

Figure 2.2: 고부담 검사와 저부담 검사의 정의

주: 정의는 AERA et al.(2014, pp. 219, 221)에서 인용.

Table 2.4: 다양한 분야에서의 검사와 평가 사용 예시

교육	고용	심리	프로그램 평가
입학	채용 전: • 능력 평가 • "무형적 요소" 평가 • 직무 미리보기 제공 • 지원자 모집	심리상태 진단	효과성 결정 및 실행
형성평가	승진	인지능력 평가	형성적 평가
학생 학습 평가	업무수행 평가	행동과 기능에 대한 통찰	비교 평가
성적 부여	법적 방어 가능성 제공	가치, 관심사 결정	프로그램 개선
미래 수행 예측		처치 계획 준비	
진단(강점, 약점)			
대학 학점			
우수상 수여			
학교/지역/국가			
모니터링			
장학금, 인턴십 수여			

이러한 이분법적 구분은 유용하지만, Tannenbaum과 Kane(2019)은 Geisinger(2011)을 따라 평가의 부담은 시험 사용과 관련된 결과에 따라 달라지며, 결과의 종류와 심각도가 다를 수 있다고 제안했습니다. 그들은 면허 시험, 취업 시험, K–12 책임성 평가와 같은 시험 적용에서 긍정적 결과와 부정적 결과, 영향, 가능성, 그리고 결과의 가역성 등 네 가지 기준을 고려할 수 있다고 주장했습니다. 예를 들어, 의사 면허 시험에서는 불합격한 지원자에게 부정적 결과가 있으며, 이는 그들이 의료 행위를 할 수 없게 되어 중요한 영향을 미칩니다. 반면, 잘못된 합격 점수는 자격이 없는 전문가를 대중에게 노출시킬 수 있습니다. 중요한 결과의 가능성은 합격 점수 근처의 응시자와 대중에게 높아지며, 결과의 지속 기간은 재시험이 허용되기까지의 시간으로, 이는 몇 달이 될 수 있습니다. 취업 선발에서는 중요성과 가능성 면에서 유사한 결과가 있을 수 있지만, 지속 기간은 덜 중요합니다. 왜냐하면 지원자는 다른 직위를 찾을 수 있기 때문입니다. 그러나 지속 기간의 또 다른 측면은 정직성 검사에서 낮은 점수를 받은 경우와 같이 시험에서 받은 피드백이 자존감에 더 오래 지속적인 영향을 미칠 수 있으며, 특히 인지된 결함을 극복하는 방법에 대한 지도가 제공되지 않을 때 그렇습니다. Tannenbaum과 Kane (2019)은 고부담 대 저부담의 이분법을 세분화한 "결과 프로필"을 제안했습니다.

타당성에 대한 위협은 평가가 고부담 또는 저부담 목적을 수행하는지, 또는 일반적으로 평가의 결과에 따라 다릅니다. 한 가지 예로, 고부담 평가에서는 부정행위가 종종 주요한 타당성 위협이 됩니다. 정의에서 언급했듯이, 평가와 관련된 부담은 반드시 시험을 치르는 사람에게만 해당되는 것이 아니라, 평가 결과에 관심이 있는 다른 사람들에게도 영향을 미칠 수 있습니다. 누가 부정행위를 할 가능성이 가장 높은지는 누가 가장 큰 이해관계를 가지고

있는지—시험 응시자, 교사, 채용 담당자, 프로그램 옹호자, 정책 결정자 등—와 관련이 있습니다. 저부담 평가에서는 동기 부여의 부족이 주요한 타당성 위협입니다(Wise & DeMars, 2005). 만약 시험 응시자가 인센티브 부족이나 다른 이유로 최적의 노력을 기울이지 않는다면, 그 시험 점수를 최적의 노력 하에서와 동일하게 해석하기는 어렵습니다. 따라서, 평가의 다양한 목적을 고려할 때 부담의 정도는 중요합니다(표 4를 참조). 고부담–저부담의 구분은 근본적으로 중요하지만 종종 간과됩니다.

2.3.1 고부담 평가의 활용

고부담 시험은 전 세계 교육 기관의 입학 여부를 결정하는 데 사용됩니다. 예를 들어, 미국의 사립 중·고등학교 입학 시험인 Secondary School Admissions Test (SSAT), 미국 대학원 입학을 위한 ETS의 GRE®, 브라질의 고등학교 졸업 인증 및 대학교 입학 시험인 Exame Nacional do Ensino Médio (ENEM), 매년 1,000만 명 이상이 응시하는 중국의 National College Entrance Examination (Gaokao), 일본 대학 입학을 위한 National Center Test, 인도의 공과대학 학부 입학 시험인 Joint Entrance Exam (JEE) 및 의대 입학 시험인 National Eligibility cum Entrance Test (NEET), 스웨덴의 Scholastic Aptitude Test (SweSAT), 호주의 Skills for Tertiary Admissions Test (STAT) 등이 있습니다. 또한, 고부담 시험은 성적 우수 장학금 지급 (예: 미국 대학 입학 시 ACT 및 SAT 성적 기반 장학금), 자격증 및 면허 시험 (예: ETS의 PRAXIS® 교사 자격 시험, 일본의 Society of Perinatal and Neonatal Medicine [JSPNM] 및 Software Testing Qualifications Board [JSTQB], 영국의 간호·조산사 면허 시험인 Objective Structured Clinical Examination [OSCE]), 채용 및 인재 선발 (예: SHL Direct, DISC Assessments, Birkman method, Predictive Index), 군사 인력 선발 및 분류 (예: 미국의 Armed Services Vocational Aptitude Battery [ASVAB], 영국 육군의 British Army Recruit Battery [BARB]) 등의 목적으로도 사용됩니다. 이 외에도, 이력서와 취업 지원서에서 경쟁력을 높이기 위해 취득하는 자격증 및 평가 기관에서 부여하는 인증(그림 1 참고)도 고부담 시험의 예시입니다. 학생들의 성적을 결정하거나 합격 여부를 판단하는 교내 시험도 고부담 시험이 될 수 있습니다. 또한, 교사나 학교에도 중요한 영향을 미칠 수 있어, 시험 대비 교육(teaching to the test)을 조장하는 요인이 될 수도 있습니다. 또한, 대학 배치 시험(placement tests)도 고부담 시험이 될 가능성이 있습니다. 2년제 또는 4년제 대학의 신입생을 대상으로 영어 및 수학 실력을 평가하는 이 시험은 학생이 대학 수업을 바로 수강할 수 있는지, 아니면 기초 과정(remedial courses)을 먼저 이수해야 하는지를 결정하는 데 사용됩니다. 다만, 일부 대학에서는 학생이 점수와 상관없이 대학 수준의 수업을 선택할 수 있기 때문에, 모든 경우에 고부담 시험이 되는 것은 아닙니다 (Bailey et al., 2010). Advanced Placement® (AP®) 시험도 마찬가지로, 성적이 좋으면 대학 학점을 인정받을 수 있다는 점에서 고부담 시험으로 간주될 수 있습니다.

한 가지 중요한 점은, 동일한 시험이 이해관계자에 따라 고부담일 수도 있고 저부담일 수도 있다는 것입니다. 예를 들어, 국가 교육 평가 시험(state accountability tests)은 학교나 교육구의 입장에서는 고부담 시험이지만, 학생 개개인에게는 저부담 시험일 수 있습니다. 이에 대한 추가적인 논의는 Tannenbaum과 Kane (2019)이 제공한 바 있습니다.

고부담 시험은 Goodhart (1984)의 법칙에 취약할 수 있습니다. Goodhart의 법칙은 특정 지표가 목표가 되는 순간, 더 이상 유효한 측정 도구로 기능하지 못한다는 원리를 말합니다. 즉, 시험의 중요성이 지나치게 강조되면, 그 시험이 본래 의도했던 평가 도구로서의 기능을

상실할 위험이 있습니다. 예를 들어, 점수 경쟁이 심화되면서 시험이 부정행위, 과도한 시험 대비 교육, 또는 시험 자체의 왜곡된 활용을 초래할 수 있습니다. 이러한 위험을 완화하기 위한 전략 중 하나는 과도한 신뢰(overconfidence)에 대한 경계입니다. 측정 과학(measurement science)은 발전을 거듭하고 있지만, 본질적으로 가정에 기반한 추정치를 생성하는 분야이므로, 시험 점수를 해석하는 대체 방식이 항상 존재할 수 있습니다 (National Research Council, 2001). 따라서, 고부담 시험의 활용과 해석에는 신중한 접근이 필요합니다.

2.3.2 고부담 사용이 혼합된 저부담 사용

저부담 검사에는 많은 종류가 있습니다. 그림 1은 지속적인 능력 향상, 학습 여정의 개별화, 새로운 영역에서의 잠재력 발견, 그리고 자신의 강점에 맞는 진로 발견을 위해 사용되는 것들을 포함한 여러 가지를 보여줍니다.

대규모 국가 교육평가(예: 미국의 국가교육발전평가[NAEP], 남아프리카의 연례국가평가[ANA])와 국제평가(예: OECD의 국제학생평가프로그램[PISA], 국제성인역량평가프로그램[PIAAC], 사회정서역량연구[SSES])는 응시자인 학생에게는 저부담 평가이며, 일부 경우에는 배경이나 맥락 설문지를 작성하는 교사, 학교, 지역에도 저부담입니다. 그러나 같은 평가가 주나 국가 정책입안자들에게는 고부담일 수 있으며, 따라서 이러한 평가의 결과는 COVID로 인한 학습 손실 발견에 대한 대응(Mervosh, 2022)이나 국가와 주가 다른 곳들과 비교하여 자신의 위치와 올바른 방향으로 가고 있는지를 볼 수 있게 하는 순위표 제시와 같은 정책적 함의를 가질 수 있습니다. 독일의 "PISA 충격"은 "뜨거운 공개 토론과 강력한 정책 대응"을 촉발했습니다(Davoli & Entorf, 2018). 결과는 환경적 효과(예: 소셜 미디어; Posso, 2016)나 세속적 추세(예: 플린 효과; Bratsberg & Rogeberg, 2018)를 평가하는 데 사용될 수 있습니다. 이러한 발견들은 응시자에게는 저부담인 평가가 정책에 상당한, 잠재적으로 의도하지 않은 결과를 미칠 수 있음을 보여줍니다(Feuer, 2012).

형성평가와 학생들의 능력 수준에 기반한 수업 맞춤화 및 피드백 제공은 평가의 또 다른 저부담 사용입니다. 적응적 교수 시스템(예: Carnegie Learning [BusinessWire, 2024], Khanmigo [DiCerbo, 2024])은 평가를 이러한 방식으로 사용합니다. 우리는 이 보고서의 피드백 섹션에서 형성평가와 피드백의 저부담 사용을 검토합니다.

또 다른 저부담 사용은 기관에 학생들의 능력에 대한, 또는 고용주의 경우 그들의 노동력에 대한 규준적 정보를 제공하는 것입니다. ETS® 주요 분야 검사는 특정 분야를 전공하는 학생들의 성취 수준에 대한 정보를 대학에 제공하기 위해 설계되었습니다. 일반적으로 해당 전공의 종합 과정에서의 데이터 수집으로 얻어지는 다양한 전공의 평가 결과는 프로그램이 교육과정을 개선하고 학생 성과를 향상시키기 위해 프로그램 효과성과 학생 수행을 평가하는 데 사용되었습니다(ETS, n.d.). 마찬가지로, OECD와 유럽연합의 교육 및 기술 온라인 프로그램은 학습자의 강점과 약점을 진단하고 국제 기준에 대비하여 훈련을 평가하기 위해 훈련생들의 문해력, 수리력, 문제해결력에 대한 정보를 제공하도록 설계되었습니다(OECD, n.d.).

상황에 따라 고부담이나 저부담이 될 수 있는 하나의 추가적인 평가 사용 사례는 AI 발전과 같은 기계 능력의 벤치마킹에서 찾을 수 있습니다. 예를 들어, PIAAC 평가는 AI 전문가들이

기계 알고리즘이 즉시 또는 예측 가능한 미래에 평가에 나타나는 문제들을 해결할 수 있는 정도를 평가하는 연구에 사용되었습니다(Elliott, 2017). 검사들은 마찬가지로 AI 챌린지 대회에서도 활용되었습니다(Friedland et al., 2004). 한 수준에서 이들은 단순히 기계 능력을 벤치마크, 이해, 진단하는 것이 목표이므로 저부담 응용입니다. 반면에, 고부담 챌린지에서 실시되는 평가는 검사를 조작하려는 일반적인 인센티브를 제공할 수 있으므로 고부담 사용으로 간주될 수 있습니다.

2.4 평가에 대한 새로운 도전과제들

검사의 다양한 용도와 그것이 제공하는 잠재적 가치에도 불구하고, 검사라는 주제는 지난 세기 동안 논란의 대상이었으며(Berman et al., 2019; Cronbach, 1975; National Research Council, 1999a, 1999b; U.S. Congress, Office of Technology Assessment, 1992) 앞으로도 그럴 것 같습니다. 여기서는 평가가 모든 사용자에게 긍정적인 결과를 제공하는 잠재력을 달성하기 위해 해결해야 할 몇 가지 새로운 도전과제들을 검토합니다.

2.4.1 검사가 충분한 가치를 제공하지 않는다는 우려

표준화 검사에 대한 불만은 오래되었습니다(Grose, 2024). 그러나 검사는 기회의 문을 열어주고 응시자, 정책입안자 및 검사 점수 정보의 다른 사용자들에게 유용한 정보를 제공할 수 있습니다. ETS(2023a)의 평가 혜택에 대한 진술에 대한 동의 수준을 특징짓는 표 5를 살펴보십시오. 응답자의 80% 이상이 평가가 직업을 찾고 배경에 관계없이 동등한 기회를 포함한 발전 기회를 제공하며, 자존감과 직업 만족도를 높이고, 새롭게 등장하는 직업과 역할에서의 능력을 측정하는 데 도움이 된다는 데 동의했습니다. 평가의 가치에 대한 이러한 긍정적인 정서는 이러한 평가 혜택에 "매우 동의"한다고 표시한 비율이 34%에서 40%에 달하는 젊은 응답자들(Z세대와 밀레니얼)에게서 특히 두드러졌습니다.

Table 2.5: 평가의 인식된 혜택

학습 평가는...	동의	매우 동의
개인이 더 나은 직업 기회와 경력 발전을 달성하는 데 도움을 준다.	85%	40%
개인의 자존감 향상에 크게 기여한다.	84%	37%
전반적인 경력 만족도 향상에 크게 기여한다.	84%	38%
발전을 위한 가치 있는 기회를 제공한다.	84%	34%
새롭게 등장하는 산업과 직무 역할과 관련된 능력을 효과적으로 측정한다.	83%	35%
능력 격차를 해소하여 발전 기회를 제공한다(예: 사회경제적, 인종적, 성별 등 서로 다른 배경에 걸쳐).	82%	34%

주: 데이터는 ETS 인간 진보 연구(ETS, 2023a)에서 가져왔음. 설문 문항: "다음 진술에 얼마나 동의하거나 동의하지 않습니까? (매우 동의하지 않음/다소 동의하지 않음/다소 동의/매우 동의)" "동의" 열은 모든 응답자 대상; "매우 동의"는 Z세대와 밀레니얼 세대만 해당; "매우 동의"는 X세대와 베이비부머 세대의 경우 약 10%-20% 더 낮음.

하지만 검사는 또한 응시자와 검사 수행 활동을 지원하는 사람들의 투자를 필요로 합니다. 이 투자는 준비와 검사 시간 및 노력, 그리고 잠재적인 평판 위험에 있습니다. 모든 관련 당사자들의 시간과 노력 지출을 정당화하기 위해 최소한 암묵적인 비용-편익 계산이 이루어집니다. 검사가 응시자와 지원자들에게 더 많은 가치를 제공할수록, 노력과 투자는 더욱 정당화될 것입니다. 따라서 검사가 응시자와 이해관계자들에게 지출을 정당화하는 검사 수익률(ROT)을 제공하는 것이 중요합니다.

검사는 종종 유용하고 실행 가능한 피드백을 제공하지 못합니다; 교육과 진로 목표를 달성하기 위한 다음 단계를 결정하는 데 도움이 될 수 있는 통찰을 사용자들에게 제공하지 못합니다. 평가의 미래는 주로 응시자와 관련된 모든 사람들에게 있어 검사의 비용-편익 비율을 변화시키기 위해 주요 이해관계자들, 특히 응시자들에게 유용한 정보를 제공하는 것과 관련될 것입니다. 검사는 현재 아는 것에서 검사가 제공하는 정보로 무엇을 할 수 있는지로 전환될 것이며, 앞으로 나아갈 방향에 대한 추천을 제공할 것입니다. 우리는 이러한 문제들을 피드백 섹션에서 다룹니다.

2.4.2 검사 초점이 너무 좁다는 우려

검사에 대한 일반적인 주장은 우리가 중요한 것을 측정하므로 검사가 우리의 가치를 나타낸다는 것입니다. 하지만 너무 자주 그 반대가 사실이 되어, 우리가 우연히 검사하고 있는 것의 중요성을 높이게 됩니다. Schrum과 Levin(2013)은 우리가 너무 자주 "모범적인 학교"의 의미를 높은 성취도 검사 점수를 산출하는 학교로 제한하여, 교육적 성취와 경제적 결과에 기여하는 훨씬 더 광범위한 능력 집합을 놓친다고 주장했습니다. 즉, 검사의 초점은 전통적으로 너무 좁았으며, 아마도 최소한 부분적으로는 가장 중요한 것보다는 측정하기 쉬운 것에 초점을 맞추었기 때문일 것입니다. 교육적 성취와 직업 및 삶의 성공은 수학과 언어 검사로 쉽게 측정할 수 있는 것을 넘어선 능력의 발달을 필요로 합니다. 평가의 미래를 위해서는 교육, 직업, 삶에 가장 중요한 능력을 식별하고 이를 평가하기 위한 타당하고 신뢰할 수 있는 방법을 개발하는 것이 중요합니다. 우리는 이러한 문제들을 이 보고서의 기술 발전의 영향 섹션에서 다룹니다.

2.4.3 점수의 타당도와 신뢰도 부족에 대한 우려

검사가 항상 측정하고자 하는 능력을 측정하는 것은 아닙니다. 예를 들어, 저부담 상황에서는 학생들이 동기부여가 되지 않고 참여하지 않을 수 있으며, 그러면 검사의 점수는 학생들이 알고 할 수 있는 것을 나타내는 유용한 지표가 되지 못합니다. 예를 들어, 우리는 대규모 평가로 주와 국가의 성취 수준을 비교하지만, 노력의 차이가 검사 점수에 영향을 미친다는 것을 알고 있음에도 불구하고, 이러한 차이에 부분적으로 책임이 있을 수 있는 노력의 차이를 고려하지 않습니다(Liu et al., 2012). 검사가 학생의 능력 수준에 대한 정확한

그림을 제공하지 못할 수 있는 또 다른 이유는 부정행위나 검사에 직접적으로 맞춘 교수를 경험했기 때문입니다. 여기서 더 일반적인 우려는 검사 과정의 보안 부족으로 인해 점수가 응시자의 능력 수준이나 학교 품질을 과대 평가하도록 허용하는 것입니다. 세 번째 불만은 특히 측정하기 어려운 능력과 구인에 대한 자기보고와 같은 약한 검사 방법과 관련이 있습니다(Stecher & Hamilton, 2014). 예를 들어, 인내심과 호기심이 학생들에게 중요한 자질일 수 있지만, 평가가 전적으로 자기보고에 의존한다면, 그러한 학생 자질에 관심이 있는 사람들이 평가에서 도출된 결론에 대한 신뢰를 잃게 할 수 있습니다. 평가의 미래는 측정하기 어려운 능력에 대한 더 나은 측정과 관련될 것 같습니다. 우리는 이러한 문제들을 이 보고서의 혁신적 측정 섹션에서 다룹니다.

2.4.4 공정성과 형평성에 대한 우려

많은 사람들이 검사에 대해 가지는 주요 우려는 검사가 모든 응시자에게 공정하고 형평성 있지 않다는 것으로, 이는 점수에 대한 전반적인 신뢰 부족을 초래하고 검사 자체에 대한 반대 태도로 이어집니다. 이러한 관점에서, 응시자가 문화, 성별, 언어, 장애 상태 또는 사회경제적 상태와 관련하여 검사 설계자와 다른 경우 검사는 능력을 정확하게 측정하지 못할 수 있습니다. 더 일반적으로, Solano-Flores(2019)는 검사가 문화적 산물이므로 검사의 타당도 논증의 일부로 다양한 문화 관련 문제들을 고려해야 한다고 주장했습니다.

또한, 검사는 공평한 경쟁의 장을 만들지 못하고 학습 기회의 차이를 반영할 수 있는 과거의 불평등을 고려하지 못하기 때문에 불공평한 것으로 볼 수 있습니다(Darling-Hammond, 2001). 결과적으로, 이러한 관점에 따르면, 검사는 그러한 불이익을 받는 사람들을 지원하지 못하고 대신 불평등, 증가하는 격차, 양극화에 기여할 수 있습니다(이러한 문제들을 다룬 Educational Assessment 특별호 소개는 Herman et al. [2023] 참조, 같은 호에서 Bennett [2023], Solano-Flores [2023], Randall [2023]의 성찰과 권고사항 참조). 이러한 검사의 실제 또는 인식된 장벽은 글로벌 맥락에서 보거나 한 국가의 학생이나 근로자가 다른 국가나 문화의 기준으로 평가될 때, 예를 들어 아시아 근로자가 미국에서 취업을 하려할 때 더욱 악화될 수 있습니다.

검사의 공정성에 관한 문제들은 교육 및 심리검사 표준(AERA et al., 2014)뿐만 아니라 ETS 품질 및 공정성 표준(ETS, 2014)과 ETS 공정한 검사 및 의사소통 개발 지침(ETS, 2022) 및 기타 유사한 문서들에서 다루어집니다. 이러한 문서의 표준이 자동적으로 실제로 이어지지는 않으며(Solano-Flores, 2023), 반드시 공정성 법적 방어에 독점적으로 사용될 법적 지원을 받는 것도 아니지만, 그럼에도 불구하고 "널리 적용 가능한 자문 출처"로 간주됩니다(Biddle & Nooren, 2006, p. 219). AERA et al.(2014, p. 2)에서 언급된 바와 같이:

:표준이 후원 기관들에 의해 강제될 수는 없지만, 이는 검사와 기타 선발 절차의 개발자와 :사용자들이 따르는 일반적으로 인정된 전문적 표준을 제시하는 것으로 규제 당국과 법원에 의해 :반복적으로 인정되어 왔습니다. 표준의 준수 여부는 사법 및 규제 절차에서 법적 책임에 대한 관련 :증거로 사용될 수 있습니다. 따라서 표준은 검사 과정의 모든 참여자들의 신중한 고려를 받을 :만합니다.

표준(AERA et al., 2014)은 공정성을 "가장 중요한 근본적 관심사"이자 "검사 개발과 사용의 모든 단계에서 주의가 필요한" "근본적인 타당도 문제"로 간주합니다(p. 49). 또한 검사 과정 동안 모든 응시자의 공정하고 형평성 있는 대우를 옹호합니다. 표준은 또한 "검사 점수의

공정하고 타당한 해석에 대한 주된 위협은 식별 가능한 응시자 집단의 점수를 체계적으로 낮추거나 높이고 의도된 용도에 대해 부적절한 점수 해석을 초래할 수 있는 검사나 검사 과정의 측면에서 비롯된다"고 주장합니다(AERA et al., p. 54). 이는 구인 무관 요소가 부적절한 검사 내용 표집, 불명확한 검사 지시문, 불필요한 문항 복잡성, 그리고 특정 집단에 유리할 수 있는 채점 기준에 의해 도입될 수 있으며 "학습 기회...가 의도된 용도에 대한 검사 점수의 공정하고 타당한 해석에 영향을 미칠 수 있다"고 제안합니다(AERA et al., 2014, p. 54).

평가의 미래에 있어 주요 도전과제는 여기서 명시된 공정성과 형평성 문제를 해결하는 것이 될 것입니다. ETS(2014, 2022)는 표준(AERA et al., 2014)에서 제기된 공정성 문제를 다루는 구체적인 지침을 제공하는 검사와 의사소통을 위한 공정성 지침을 개발했습니다. ETS(2022)는 네 가지 기본 원칙을 제시했습니다: (a) 의도된 구인의 중요한 측면을 측정한다; (b) 응시자의 성공에 대한 구인 무관 장벽을 피한다; (c) 다양한 응시자들이 알고 있는 것과 할 수 있는 것을 보여줄 수 있도록 하여 타당한 추론이 지지되는 평가 설계, 내용, 조건을 제공한다; (d) 다양한 응시자 집단에 대한 타당한 추론을 지지하는 점수를 제공한다. ETS(2022)는 이러한 일반 원칙을 지원하는 구체적인 지침을 이어서 제시했습니다.

검사 공정성에 대한 우려 외에도, 형평성에 대한 우려가 있습니다. 집단 간 검사 수행의 격차는 최소한 부분적으로 학습 기회의 차이를 반영할 수 있으며, 검사는 기회 격차를 식별하는 데 도움이 될 수 있습니다(National Academies of Science, Engineering, and Medicine, 2019). 그러나 검사가 불의를 전파한다는 견해는 검사가 배경에 관계없이 학생들의 학업 성취를 반영하며, 입학에서 고려되는 성적과 다른 측정치보다 더 높은 예측 정확도를 가지므로 저소득층과 소수 집단 지원자들에게 기회를 제공한다는 재활성화된 견해에 의해 도전받고 있습니다(Deming, 2024; Flanagan, 2021; Leonhardt, 2024; McWhorter, 2024). 더욱이, 검사는 교수의 한 형태로 기능할 수 있으며 이를 통해 형평성 문제를 다룰 수 있습니다; 우리는 평가의 미래에서 주요 초점이 "공평한 학습 기회를 다루는" 교육을 위한 평가를 달성하는 방법을 개발하는 것이 될 것으로 예상합니다(The Gordon Commission, 2013, p. 150). 우리는 검사가 피드백을 제공하는 문제를 피드백 섹션에서 다룹니다.

2.5 평가의 미래 전망

이 논문에서, 우리는 앞 하위 섹션에서 확인된 도전과제들과 우려사항들을 다루고, 평가의 미래를 위한 포괄적인 주제는 평가가 능력 기반이 되고, 기술이 향상되며, AI와 관련 기술의 발전에 의해 주도될 것이라고 주장합니다. 학습 증진에서의 역할을 인정하여, 미래의 평가는 부족함에 덜 초점을 맞추고, 학습자들이 교육과 진로 목표를 달성하는 데 도움이 되도록 그들의 강점을 바탕으로 발전하도록 안내할 것입니다. 부정적인 피드백은 특히 자원을 통제할 능력이 적은, 저영향력 개인들의 동기와 수행 수준에 해로운 것으로 나타났습니다(Straub et al., 2023). 미래의 평가는 응시자 중심이며 응시자가 취할 수 있는 구체적인 행동에 초점을 맞춘, 실행 가능한 피드백을 제공할 것입니다.

Table 2.6: AI가 능력의 미래에 미치는 영향

예측	동의 + 매우 동의
직장에서의 AI로 인해, 대부분의 직원들은 자신의 능력을 습득하거나 갱신할 필요가 있을 것이다.	85%
AI는 근로자들이 기술적 능력과 인간적 능력의 조합을 갖출 것을 요구할 것이라고 생각한다.	83%
AI는 직장에서 필수적인 기술의 재평가를 촉진할 것이다.	83%
AI는 경력 전환, 재숙련화, 자기 재창조의 필요성을 증폭시킬 것이다.	80%
AI가 오늘날 존재하지 않는 새로운 직업 기회를 창출할 것이라고 믿는다.	72%

주: 데이터는 ETS 인간 진보 연구(ETS, 2023a)에서 가져왔음. 설문 문항: "다음 진술에 얼마나 동의하거나 동의하지 않습니까? (매우 동의하지 않음/다소 동의하지 않음/다소 동의/매우 동의)" AI = 인공지능

2.5.1 응시자와 이해관계자에게 유용한 정보 제공

미래의 평가는 응시자와 다른 이해관계자들에게 유용하고, 이해하기 쉬우며, 신뢰할 수 있고, 타당하며, 공정하고, 신뢰할 만한(안전한 과정에 기반한) 정보를 제공하도록 노력해야 합니다. 평가는 비용 효과적이어야 하고, 관련된 언어로 제공되어야 하며, 가능한 경우 통찰을 도출하거나 실행 가능해야 합니다. 그 정보는 응시자가 추가 교육과 현재 및 미래 직업을 위해 가장 중요한 능력에서 어디에 서 있는지를 보여주는 자격증, 점수, 배지 및 기타 지표의 형태를 취할 수 있으며, 개별 응시자가 교육과 진로 목표를 달성할 수 있는 방법에 대해 응시자와 이해관계자들에게 정보를 제공하는 실행 가능한 피드백이 함께 제공됩니다.

2.5.2 핵심 능력의 식별

응시자에게 유용한 정보를 제공하기 위해서는 교육과 진로 목표를 달성하는 데 필요한 가장 중요한 능력을 식별하는 것이 필요합니다. 핵심 능력을 식별하기 위해서는 어떤 능력이 중요성이 증가하고 어떤 능력이 쓸모없게 될 것인지를 결정하는 데 도움이 되는 다양한 방법론—설문조사, 직업 동향, 재정 스캔—을 사용하여 능력의 미래 생존 가능성에 대한 증거를 수집해야 할 것입니다. 이러한 분석을 수행하는 것(예: Autor et al., 2024; Eloundou et al., 2023; Frey & Osborne, 2017; Lassébie & Quintini, 2022)은 학교, 직업, 사회를 위해 어떤 능력에 투자할지 결정하는 투자 결정을 돕는 지표를 생산할 수 있게 할 것입니다.

2.5.3 측정하기 어려운 능력 평가를 위한 방법 발전

의사소통, 창의성, 협력과 같이 오늘날 점점 더 중요해지고 미래에 중요성이 더욱 커질 것 같은 많은 능력들은 측정하기 어려운 능력입니다(표 6 참조). 이들은 측정하기 어렵지만 중요하기 때문에, 우리는 이들을 측정하기 위해 단순한 자기보고와 타인 평가를 사용하는 경향이 있습니다. 하지만 이러한 방법들은 수학과 읽기와 같은 기술적 능력, 즉 이른바 하드 스킬을 측정하는 데 사용하는 방법만큼 강력하지 않습니다. 자기보고와 타인 보고는 계속 사용될 것이지만, 이들은 반응 양식(측정되는 구인에 관계없이 비슷한 방식으로 반응하는 경향; He et al., 2014), 후광(대상이 측정되는 속성에 관계없이 대상을 같은 방식으로 평가하는 경향; Cooper, 1981), 그리고 준거 편향(응답자가 평가에서 다른 기준을 사용하는 경향; Lira et al., 2022)과 같은 잘 문서화된 편향들과 관련이 있습니다. 이러한 측정을 보완하거나 대체하기 위해, 주관적 평정에 의존하지 않는 게임, 시뮬레이션, 상호작용 및 협력 과제를 포함한 매력적이고, 개별화되며, 맥락화된 수행 과제를 개발할 필요가 있습니다. 평가의 미래를 위한 하나의 경향은 지나치게 표준화된 접근에서 벗어나 "개별화되고, 차별화되며, 적응적이고, 문화적 언어적으로 관련되며, 맥락 기반적"인 것으로 더 잘 특징지어질 수 있는 접근으로 나아가는 것일 수 있습니다(Morell, 2017, p. 2). Sireci(2020)는 검사 조건과 상호작용할 수 있는 개인적 특성을 이해하고 그러한 개인적 특성을 수용하는 것이 "학생들의 진정한 능숙도에 대한 더 정확한 해석"으로 이어질 잠재력이 있다고 주장했습니다(p. 101).

평가의 미래는 또한 키스트로크, 대화, 반응 시간, 그리고 발달 과정과 능력의 상태에 대한 추론을 도출하는 데 사용될 수 있는 기타 학습 및 수행 지표를 포함한 프로세스 데이터의 분석을 포함하여, 자연적으로 발생하는 행동을 측정하는 방법의 개발을 포함할 것입니다. 이러한 방법들은 정의적, 행동적, 또는 인지적(ABC; Liu, Kell, et al., 2023) 어떤 유형의 능력에도 적용될 수 있다는 점에 주목하십시오.

이 노력의 중요한 부분은 우리가 새로운 측정을 고안하는 데 얼마나 성공적이었는지를 평가하는 지표를 개발하는 것이 될 것입니다. 우리는 타당도, 신뢰도, 공정성과 형평성을 포함한 전통적인 심리측정 지표에 의존할 수 있습니다. 또한 우리는 우리의 노력이 얼마나 가치 있게 여겨지고 핵심적이고 확장되는 시장과 얼마나 부합하는지를 통해 성공을 평가할 수 있습니다.

2.5.4 개별화된 피드백을 통해 응시자와 다른 이해관계자들에게 기회 제공

응시자에게 유용한 피드백을 제공하기 위해서는 효과성 증거와 함께 여러 분야에서 나오는 학습 원리의 식별과 실행이 필요할 것입니다. 이러한 분야들은 교육심리학, 인지심리학, 산업-조직심리학, 학습과학, 신경과학을 포함합니다. 인적 요인, 훈련, 인간-컴퓨터 상호작용과 같은 응용 영역과 컴퓨터 지원 협력 학습과 적응적 학습 또는 지능형 교수 시스템과 같은 교수 영역도 검사 실제에 통합될 수 있는 발견과 원리를 제공할 수 있습니다. AI와 교육 분야의 중요한 연구(Koedinger et al., 2023; Zapata-Rivera & Hu, 2022)는 검사가 어떻게 응시자들에게 그들의 학습을 향상시키고 검사로부터 받는 혜택을 증진시키기 위한 유용한 정보를 제공할 수 있는지에 대해 알려줄 수 있습니다. 피드백 제공은 또한 지속적인 과정이어야 합니다. ETS(2023a)의 응답자 87%가 "학습 평가는 수행의 일회성 스냅샷이

아닌 지속적인 피드백을 제공해야 한다"는 데 동의했습니다. 피드백의 혜택은 응시자에게만 국한되어서는 안 됩니다—정책입안자, 교사 및 다른 이해관계자들도 유익하고 실행 가능한 피드백으로부터 혜택을 받을 수 있습니다.

피드백 제공은 모든 종류의 교육 및 건강 개입과 유사성을 공유하므로 그러한 분야들로부터 교훈을 얻을 수 있습니다. 예를 들어, "연구 결과와 다른 증거 기반 실천을 일상적 실천으로 체계적으로 받아들이는 것을 촉진하는 방법의 과학적 연구, 따라서 건강 서비스의 질과 효과성을 향상시키는 것"으로 설명되는 더 넓은 실행 과학 분야(Bauer et al., 2015)는 피드백 관리가 어떻게 학습자 성과를 향상시킬 수 있는지에 대한 유용한 지침을 제공할 수 있습니다. 교훈은 또한 실행을 통한 학습을 가속화하고, 새로운 도구와 과정의 개발과 개선을 안내하도록 설계된 개선 과학으로부터도 올 수 있습니다(Hinnant-Crawford, 2020).

2.5.5 평가의 미래를 위한 주제와 논문의 구성

우리는 다음 네 섹션에 걸쳐 다룰 주제들을 중심으로 평가의 미래를 조직하는 것이 유용하다고 믿습니다. 주제들은 종종 겹치지 않는 문헌에서 다루어지는, 구별되는 작업 본체와 과학적 배경을 반영합니다. 그러나 주제들에 나타난 모든 전선에서의 발전은 평가의 미래에 필수적입니다. 다음 섹션인 '미래를 위한 능력: 기술 발전의 영향'은 주로 경제학과 AI 연구에 기반합니다. '혁신적 측정: 측정하기 어려운 능력을 평가하기 위한 새로운 접근'은 인지심리학, 산업-조직심리학, 성격심리학 등 다양한 분야에서 가져옵니다. 'AI와 기술 중심 발전을 통한 운영 혁신' 섹션은 주로 검사 개발, 채점, 보고에 있어 검사 산업의 전통적 관심사를 반영하며 교육측정과 심리측정, 운영연구, AI 등에서 가져옵니다. '피드백: 학습과학 주도의 통찰과 응시자를 위한 실행 계획' 섹션은 인지심리학, 교육심리학, 학습과학, 적응적 교수에서 가져옵니다. 우리는 '요약과 결론' 섹션으로 마무리합니다.

3 미래를 위한 능력: 기술 발전의 영향

"현재 가장 큰 불일치는 능력의 질과 관련성에 있다." Andreas Schleicher, OECD 교육 및 능력 국장

우리는 평가의 미래가 주로 생산적이고 정보를 갖춘 시민이 되고, 건강과 웰빙을 유지하며, 공동체와 사회에 기여하는 구성원이 되는 데 필요한 능력에 의해 주도될 것이라고 믿습니다. 기업가들은 그러한 능력을 원할 것이고, 고용주들은 새로운 근로자를 고용할 때 그러한 능력을 찾고 현재 인력을 위해 그러한 능력을 개발할 것입니다. 고등교육은 그들의 제공 과정과 전공 및 기타 자격증과 인정 형태의 인기도에서 능력에 대한 수요에 대응할 것입니다. K-12 교육도 이를 따라 표준과 교육과정을 발전시킬 것입니다—지난 10년간 사회정서학습(SEL) 능력에 대한 표준, 교육과정, 평가의 성장이 좋은 예입니다(Burrus et al., 2022). 정부와 산업계는 그러한 능력의 개발을 보장하기 위한 연구개발 의제를 개발할 것입니다. 모든 경우에, 교육자와 고용주는 입학, 채용, 승진, 학생과 인력 개발에서 좋은 결정을 내릴 수 있도록 학생, 지원자, 재직자의 능력에 대해 알고 싶어할 것입니다. 이것이 평가의 역할이었고 앞으로도 그럴 것입니다.

오늘날 다를 수 있는 것은 기술과 AI의 발전으로 인한 빠른 변화의 속도입니다. 아직 예견되지 않은 능력을 요구하게 될 경제와 미래 직업 및 직종의 본질의 전면적 개편을 예측하는 미래 전망이 부족하지 않습니다. Dell(2018)의 최근 연구에 따르면, 설문에 응한 3,800명의 글로벌 비즈니스 리더 중 56%가 "학교는 아직 존재하지 않는 직업을 위해 학생들을 준비시키기 위해 무엇을 배워야 하는지가 아닌 어떻게 배워야 하는지를 가르쳐야 할 것"이라고 추측했습니다. ETS 인간 진보 연구(ETS, 2023a)에서, 응답자들은 AI가 능력을 갱신하고, 기술적 능력과 인간적 능력을 결합하며, 아직 존재하지 않는 새로운 직업 기회를 위한 능력을 식별하는 필요성에 영향을 미칠 것 같은 많은 영역을 지적했습니다(표 6). Avalanche VC의 파트너인 Eric Lavin이 ETS 인간 진보 연구에서 언급했듯이, "배우는 방법을 배우는 것이 아마도 핵심 능력일 것입니다. 더 많은 기술이 들어오면서 능력의 반감기가 짧아지고 있습니다. 가장 중요한 능력은 인간이 되는 것과 해야 할 일과 공명하는 방식으로 새로운 도구를 사용하는 방법을 배우는 것입니다."

이 섹션은 다음과 같이 구성됩니다. 첫째, 우리는 ETS 인간 진보 연구(ETS, 2023a)의 응답자 인식, 고용주와 교육자 설문, 구인 광고 분석, 연구, 정책, 실천의 동향에 기반하여 오늘날 수요가 있는 능력을 검토합니다. 다음으로, 우리는 오늘날 가장 많이 찾는 능력이 지난 100년간 평가의 관심 초점이었던 전통적인 교육과정 능력이 아니라, 대신 평가의 도전과제를 제시하는 측정하기 어려운 능력이라는 점에 주목합니다. 다음으로, 우리는 동향 분석과 AI와 새로운 기술이 능력의 변화하는 본질에 미치는 영향 분석에 기반하여 미래에 수요가 있을 것 같은 능력을 논의합니다. 우리는 평가의 미래에 대한 함의에 대한 논의로 마무리합니다.

3.1 오늘날 요구되는 능력

우리는 세 부문에서 능력의 중요성에 대한 증거를 검토합니다: 직업, 고등교육, K-12 교육입니다. 서로 다른 부문에서 요구되고 개발되는 능력은 다를 수 있으며, 이러한 능력을 식별하는 방법론도 부문별로 다릅니다. 그림 3은 ETS 인간 진보 연구(ETS, 2023a)의 응답자들이 직업 시장이나 삶의 성공을 위해 필요하다고 생각한다고 표시한 능력을 보여줍니다. 직업 시장의 경우, 기술적 능력이 선두였고, 그 뒤를 창의성, 의사소통, 디지털 문해력이 따랐습니다. 삶의 성공을 위해서는 의사소통과 문제해결이 선두였고, 그 뒤를 창의성, 기술적 능력, 시간 관리, 인내심이 따랐습니다.

3.2 고용주들이 찾는 능력

새로운 채용에서 고용주들이 찾고 있고 현재 인력에서 개발하고자 하는 능력의 종류를 결정하기 위해 다양한 접근이 시도되었습니다. 고용주 설문은 고용주들이 직원들이 가져야 한다고 말하는 능력이 무엇인지를 반영합니다; 구인 광고 분석은 고용주들이 현재 채용하고 있는 능력을 식별하며, 이는 설문 응답과 일치해야 하지만 반드시 그럴 필요는 없습니다. 고용주 설문과 구인 광고를 모두 검토하는 것이 유용합니다. 표 7은 이러한 접근들을 사용한 여러 연구의 결과를 제시합니다. 이 연구들은 표 다음의 섹션들에서 요약됩니다.

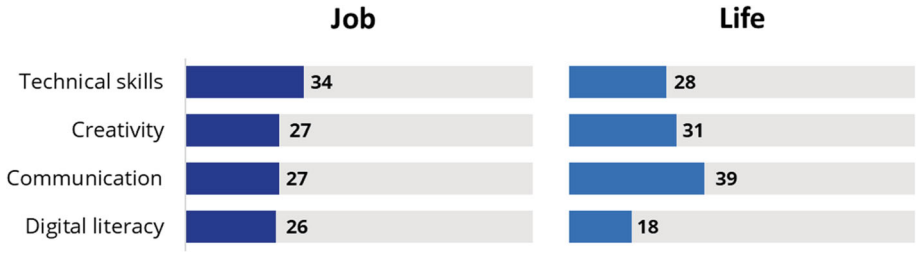

Figure 3.1: 직업 시장 또는 삶의 성공을 위해 필요한 다양한 능력 선택 비율

주: ETS 인간 진보 연구(2023a)에서 가져옴. 설문 문항: '향후 2-3년 동안 직업 시장에서 경쟁력을 갖추기 위해 어떤 능력을 습득(또는 향상)해야 한다고 생각하십니까? 최대 3개까지 선택하세요.' 11%가 해당 없음 선택. '삶의 성공을 위해 가장 필요한 능력은 무엇입니까? 최대 3개까지 선택하세요.' 3%가 해당 없음 선택.

Table 3.1: 고용주 설문과 구인 광고 분석에 기반한 최고 평가 능력

고용주 설문에서 가장 높은 평가를 받은 능력		
NACE (2022)	**Wilkie (2023)**	**WEF (2021)**
문제해결 능력	의사소통 능력	분석적 사고
팀워크 능력	경청 능력	창의적 사고

고용주 설문에서 가장 높은 평가를 받은 능력			
강한 직업윤리		비판적 사고 능력	회복력, 유연성, 민첩성
분석적 및 정량적 능력		대인관계 능력	동기부여와 자기인식
의사소통 능력			
기술적 능력			

구인 광고 분석에 기반한 가장 높은 평가를 받은 능력			
Rios et al. (2020)	**Shafer et al. (2023) 지구과학자**	**Mankki (2023) 교사 교육**	**Burning Glass Technologies (2019) 기초 능력**
구두 및 문서 의사소통	문서 의사소통	대인관계 능력, 팀워크	의사소통 능력
협력		문화적 다양성에 대한 민감성	팀워크와 협력
문제해결		문화간 이해	조직 능력
의사소통		전문성	문제해결
사회적 지능		리더십 능력	
자기주도성			

주: NACE = 전국대학취업협회; WEF = 세계경제포럼. Burning Glass는 능력을 기술적 능력, 소프트웨어 능력, 142개의 기초 능력의 세 범주로 나누었음. 여기서는 기초 능력의 요약만 보고됨.

표 7은 경청 능력을 포함한 의사소통 능력이 연구들에 걸쳐 높은 평가를 받고 있음을 보여주며, 비판적 사고, 분석적 사고, 기술적 능력, 문제해결과 같은 인지적 능력 군집; 팀워크, 협력, 사회적 지능, 사회적 기술과 같은 대인관계 능력; 직업윤리, 조직 능력, 자기주도성, 동기부여, 자기인식과 같은 개인내적 능력도 마찬가지입니다. 응답자들은 또한 문화적 능력(문화적 다양성에 대한 민감성; 문화간 이해)을 언급했습니다. 용어의 변화를 고려할 때 서로 다른 연구들에 걸쳐 발견된 결과들 간에 상당한 중첩이 관찰되는 것도 주목할 만합니다. 이러한 목록들은 능력의 미래에서 중요한 모든 것을 반영하지는 않는데, 특히 직업의 관점에 초점을 맞추고 있고 중요한 생활 능력을 반영하지 않기 때문입니다; OECD(2015)는 건강, 가정생활, 시민 참여(OECD, 2023), 그리고 삶의 만족이 개인의 웰빙과 사회적 진보에 미치는 중요성을 주장했습니다. 또한 이러한 목록들은 수학, 언어, 과학 능력과 같이 교육과정의 일부로 정규 교육에서 가르치는 중요한 지식을 포착하지 않습니다. 그럼에도 불구하고, 위의 목록들은 아직 평가가 완전히 개발되지 않은, 따라서 미래 성장을 위한 기회를 나타내는 중요한 능력 집합을 제공합니다. 다음 두 섹션에서 이러한 능력의 요약을 더 깊이 있게 논의합니다.

고용주 설문에서 찾는 능력. 고용주 설문은 기업 리더들에게 지원자의 이력서에서 무엇을 찾는지, 어떤 능력이 중요하다고 생각하는지, 그리고 관련 주제들에 대해 조사합니다. 이러한 설문의 한계를 인정하는 것이 중요합니다: 표본이 무작위가 아니고, 종종 작으며, 응답 편향의 영향을 받을 수 있고, 질문의 표현이 답변에 영향을 미칠 수 있습니다. 그럼에도 불구하고, 고용주 설문은 직업 현장에서 다양한 능력에 대한 고용주 선호도에 대한 일부 증거를 제공합니다.

전국대학취업협회(NACE)는 지난 10년 정도 동안 미국 고용주들을 대상으로 연례 설문을 실시해 왔습니다. 최근의 취업 전망 보고서(NACE, 2022)에서 그들은 고용주들이 이력서에서 찾는 최상위 속성이 문제해결 능력, 팀워크 능력, 강한 직업윤리, 분석적 및 정량적 능력, 의사소통 능력, 기술적 능력이었으며, 50%에서 61%의 고용주들이 이러한 능력을 매우 또는 극히 중요하다고 평가했음을 발견했습니다. 또한 학점으로 지원자를 선별하는 고용주의 비율이 지난 4년 동안 73%에서 37%로 급격히 감소했음을 발견했는데, 이는 아마도 교육과정 성취에 대한 고용주의 관심이 감소하고 때로는 지속가능한 능력이라고 불리는 것에 대한 관심이 증가했음을 나타낼 수 있습니다. 650명의 고용주를 대상으로 한 Cengage/Morning Consult 설문(Cengage, 2019)은 수요가 가장 높은 능력이 의사소통 능력, 경청 능력, 비판적 사고 능력, 대인관계 능력이었으며, 이러한 능력이 그들의 조직에서 리더십 직위를 얻는 데 매우 중요하다고 말한 고용주의 비율이 73%에서 77%에 달했음을 발견했습니다. 이 두 연구의 결과는 본질적으로 15년 전의 연구(Casner-Lotto & Barrington, 2006) 결과를 복제한 것인데, 그 연구는 고용주들에게 다양한 능력이 성공에 얼마나 중요한지 물었고 전문성/직업윤리, 팀워크/협력, 비판적 사고/문제해결, 구두 및 문서 의사소통, 다양성, 리더십을 포함하는 응용 능력이 수학, 읽기, 과학과 같은 기초 지식/능력보다 매우 중요하다고 평가될 가능성이 더 높다는 것을 발견했습니다.

미국 설문을 보완하기 위해, 세계경제포럼의 설문(Di Battista et al., 2023)은 미국 외 고용주들을 연구하며 오늘날의 핵심 능력에 대해 물었습니다. 고용주들은 분석적 사고; 창의적 사고; 회복력, 유연성, 민첩성; 동기부여와 자기인식; 호기심과 평생학습; 기술 문해력; 신뢰성과 세부사항에 대한 주의; 공감과 적극적 경청; 리더십과 사회적 영향력을 가장 높이 평가된 능력으로 식별했으며, 응답자의 39%에서 67%가 이들을 핵심 능력으로 평가했습니다. 고용주들은 또한 미래(지금부터 5년 후)에 대해서도 질문을 받았고 분석적 사고와 창의적 사고가 여전히 상위를 차지했지만, 호기심과 평생학습 그리고 기술 문해력이 상당히 증가하여 회복력, 유연성, 민첩성과 함께 미래를 위한 예측된 상위 5개 능력 군집에 합류했습니다.

구인 광고에서 찾는 능력. 고용주들이 인력 진입자를 모집하는 수단인 구인 광고를 분석하는 것은 직업 현장에서 다양한 능력의 가치를 결정하는 데 도움이 되는 고용주 설문을 보완해야 합니다. Rios et al.(2020)은 구인 목록을 위한 취업 웹사이트에서 142,000개의 구인 광고를 조사했고 구인 광고의 70%가 "21세기 능력"을 요구했음을 발견했습니다. 가장 많이 요구된 능력은 구두 및 문서 의사소통, 협력, 문제해결, 의사소통 능력, 사회적 지능, 자기주도성이었습니다. Shafer et al.(2023)도 마찬가지로 문서 의사소통이 학사 수준 지구과학자에게 가장 자주 요구되는 능력(67%)임을 발견했습니다. Mankki(2023)는 교사 교육자 일자리 광고를 요약하며 가장 자주 열거된 개인적 자질이 대인관계 능력과 팀워크, 문화적 다양성과 문화간 이해에 대한 민감성, 전문성, 리더십 능력이었음을 발견했습니다.

Burning Glass Technologies(2019, p. 14)는 구인 광고 분석을 수행하여 찾는 능력을 기술적 능력, 소프트웨어 능력, 142개의 기초 능력의 세 범주로 나누었습니다. 18개 경력 분야 전체에 걸쳐 최상위 기초 능력은 의사소통 능력이었습니다; 팀워크와 협력은 14개 분야에 걸쳐 상위 5위 안에 평가되었습니다; 조직 능력은 18개 분야 중 17개 분야에서 상위 10위 안에 평가되었으며, 문제해결은 특히 고객 지원과 공학 분야에서 매우 높이 평가되었습니다.

3.2.1 고등교육에서의 능력

최근 Forbes 잡지 기사는 "소프트 스킬 논쟁이 끝났다"고 선언했습니다(Flynn, 2023). Flynn은 1991년 영향력 있는 필수 능력 달성을 위한 장관 위원회(SCANS) 보고서 발간 이후, 기초 능력(읽기, 쓰기, 수학, 말하기, 듣기), 사고 능력(창의성, 의사결정, 문제해결, 마인드 아이, 학습 방법 알기, 추론), 그리고 개인적 자질(책임감, 자존감, 사교성, 자기관리, 청렴성)의 기초의 중요성이 잘 문서화되어 왔음을 지적했습니다. 그러나 고용주들은 대학 졸업생들이 점점 더 가치를 인정받고 있는 중요한 소프트 스킬이나 지속가능한 능력을 갖추지 못했다고 한탄합니다(Wilkie, 2023).

ETS는 지난 20년 동안 대학 관리자와 교수진에게 고등교육에 진입하는 학생들에게 중요하다고 생각하는 자질과 고등교육 과정에서 개발하는 것이 중요한 자질이 무엇인지를 물어보는 여러 연구를 수행했습니다. 이러한 연구들은 인터뷰, 포커스 그룹, 설문조사 등 다양한 접근법을 사용했습니다. 가장 자주 지명된 속성들 중에는 인내(끈기, 회복력, 추진력, 직업윤리), 전문성(조직력, 시간 관리, 자기 규율, 신뢰성, 믿음성), 동기부여, 해당 분야에 대한 열정과 관련된 것들이 있습니다. Oswald et al.(2004)은 대학 사명 선언문을 분석하여 지적 행동(지식, 학습, 일반 원리의 숙달; 지속적 학습과 지적 관심 및 호기심; 예술적 감상과 호기심), 대인관계 행동(다문화 관용과 감상; 대인관계 능력; 사회적 책임, 시민의식, 참여), 그리고 개인내적 행동(신체적 및 심리적 건강; 경력 지향성; 적응성과 생활 능력; 인내; 윤리와 청렴성)으로 군집화되는 대학 수행의 12가지 차원을 식별했습니다.

국가과학공학의학원(Hilton & Herman, 2017)은 이러한 능력들이 고등교육 과정의 지속과 성공에 관련되어 있고 개입을 통해 향상될 수 있다는 증거를 바탕으로 중요한 대인관계 및 개인내적 능력을 식별하고자 했습니다. 그들은 8가지를 식별했습니다: 성실성 행동, 소속감, 학업적 자기효능감, 성장 마인드셋, 효용 목표와 가치, 내재적 목표와 흥미, 친사회적 목표와 가치, 그리고 "긍정적 미래 자아"입니다. 그러나 이 보고서는 또한 이러한 능력의 측정이 질이 낮다고 지적했습니다: 거의 전적으로 자기보고에 의존하고 평가의 심리측정적 특성—신뢰도, 타당도, 공정성—이 언급되는 경우가 드물었습니다. 이러한 발견은 평가의 미래를 위한 명확한 기회를 나타냅니다.

3.2.2 K-12에서 중요한 능력

K-12에서 사회정서 학습 능력을 평가하는 첫 대규모 노력 중 하나는 캘리포니아 교육개혁청(CORE) 프로젝트였으며, 이는 낙오학생방지법(NCLB) 법안의 면제를 받아 수행되었습니다. 현재 발견과 교훈에 대한 많은 보고서가 있지만(Krachman et al., 2016; Meyer et al., 2018; West et al., 2018), 이 섹션에서 중요한 것은 프로젝트가 핵심 능력을

어떻게 식별했고 어떤 결론을 내렸는지입니다. Krachman et al.(2016)은 사회정서 학습 전문가와 CORE 지역구 대표들이 참여한 2013년의 초기 회의를 회고했습니다. 전문가 직원들은 "의미 있고, 측정 가능하며, 변화 가능한" 주제를 제안했습니다. 지역구 직원들은 최소한 하나의 대인관계 요인과 하나의 개인내적 요인을 식별하는 것을 우선시했습니다. 투표 과정 후, 성장 마인드셋, 자기효능감, 자기관리, 사회적 인식이라는 네 가지 역량이 도출되었으며, 다섯 번째인 협력적 문제해결은 거의 포함되었지만 PISA 2015의 협력적 문제해결에 대한 결과를 기다리기 위해 보류되었습니다. (이러한 차원들의 식별은 어느 정도 학업적, 사회적, 정서적 학습을 위한 협력체[CASEL]의 영향을 반영하며, 이는 다시 발달심리학과 사회심리학의 영향을 반영합니다.)

한편, 여러 연구들이 교육에서 Big 5 성격 요인(성실성, 친화성, 외향성, 정서적 안정성, 개방성; Mammadov, 2022)의 중요성에 대한 증거를 발견했고, 이는 OECD가 SSES에서 그 모델을 채택하게 했습니다(Chernyshenko et al., 2018; OECD, 2021). Big 5는 산업계와 군대 입대 또는 인사 선발과 분류 검사에서 널리 사용되는 성격심리학 모델입니다. 결국, 두 프레임워크는 그들의 서로 다른 기원에도 불구하고 근본적으로 그렇게 구별되지 않습니다(Soto et al., 2022). OECD(2023)는 첫 연구에서 사용된 등급척도 측정의 대안을 추구하며 이 연구의 후속 연구를 계획하고 있습니다.

"졸업생의 초상"은 여러 주에서 채택하고 주 교육위원회 전국협회(Norville, 2022)가 지원하는 프레임워크로, 주들이 "학생들이 고등학교를 졸업하기 전에 숙달해야 할 능력과 지식을 더 잘 정의할 수 있게" 합니다(p. 1). 이는 역량 기반 교육 접근법을 채택하고(Patrick, 2021) 의사소통과 비판적 추론과 같은 강조 영역을 결정하기 위해 이해관계자들과 협력하여 프로필을 정의하는 것을 포함합니다. 예를 들어, 사우스캐롤라이나의 졸업생 역량 프레임워크는 12가지를 제안합니다: 비판적으로 읽기, 아이디어 표현하기, 탐구를 통해 조사하기, 정량적으로 추론하기, 출처 사용하기, 해결책 설계하기, 독립적으로 학습하기, 갈등 다루기, 팀 이끌기, 네트워크 구축하기, 웰니스 유지하기, 시민으로서 참여하기. 다른 주들도 비슷한 아이디어를 개발했거나 추구하고 있습니다.

다른 능력의 가치를 식별하는 또 다른 접근법은 OECD의 PISA 프로그램의 특별 주제 영역을 고려하는 것입니다. PISA는 2000년부터 시작하여 3년마다 실시되며, 매 주기마다 15세 학생들의 읽기, 수학, 과학을 검사합니다; 가장 최근 조사에는 81개국이 참여했습니다(OECD, 2022b). 또한, PISA는 네 번째 혁신 영역 평가를 추가하여 변화하는 교과 횡단적 역량을 측정합니다. 혁신 영역 평가를 식별하는 과정은 참여 국가들과의 협상을 포함하므로, 식별된 주제는 전 세계적으로 해당 주제의 인기도를 간접적으로 가늠할 수 있습니다. 2012년 이후, PISA의 혁신 영역 평가는 문제해결, 금융 문해력, 협력적 문제해결, 글로벌 역량, 창의성에 대해 이루어졌습니다.

OECD의 2030년을 위한 능력: 개념적 학습 프레임워크(2019)는 OECD의 교육과 능력의 미래 2030 프로젝트에 참여한 국제 이해관계자 그룹의 의견을 바탕으로 했습니다. 이 보고서는 능력을 "과정을 수행하고 목표를 달성하기 위해 자신의 지식을 책임 있게 사용할 수 있는 능력과 역량"으로, 그리고 "복잡한 요구를 충족시키기 위해 지식, 능력, 태도, 가치를 동원하는 것을 포함하는 전체적인 역량 개념의 일부"로 정의했습니다(OECD, 2019, p. 4). 이 그룹은 세 영역의 능력을 우선시했습니다: 인지적 및 메타인지적 능력(비판적 사고, 창의적 사고, 학습하는 법 배우기, 자기조절); 사회적 및 정서적 능력(공감, 자기효능감, 책임감, 협력); 실천적 및 신체적 능력(새로운 정보통신기술 기기 사용하기). 또한 그들은 지식과 태도 및

가치가 서로 얽혀 있으며 지식과 능력을 발달시키는 데 필수적이라고 지적했습니다. 그들은 인지적 능력이 복잡한 문제를 해결하고 AI와 상호보완적인 방식으로 일하는 데 필수적이라고 주장했습니다. 창의성은 계속 유효할 것 같으며, 문제해결과 비판적 사고와 같은 고차원적 능력은 계속 중요할 것입니다. 메타인지적 능력은 평생학습의 핵심이며, 이는 AI 발전과 함께 점점 더 중요해질 것입니다. 문화적 이해와 불확실성 다루기도 기술 발전이 가져오는 변화에 적응하는 데 핵심입니다. 사회적 및 정서적 능력은 이제 필수적인 것으로 인식되고 있으며 인구통계학적 및 사회적 변화와 함께 계속 그럴 것입니다; AI 또한 사회적 및 정서적 능력이 필요한 직업에서 근로자를 대체할 가능성이 낮습니다. 예술과 건강한 습관 및 운동 루틴의 발달을 포함하는 실천적 및 신체적 능력은 건강과 웰빙을 지원함으로써 개인에게 계속해서 이익이 될 것입니다.

3.3 수요가 있는 능력의 측정 난점

이 섹션의 이전 하위 섹션들은 학습하는 법 배우기, 창의성, 의사소통, 비판적 사고, 문화적 이해, 호기심, 유연성, 회복력 등 다양한 능력의 평가에 대한 사례를 개괄했는데, 이들은 아직 어떻게 가장 잘 측정될 수 있는지에 대한 명확한 합의가 없는 능력들입니다. 국가과학공학의학원(2018) 보고서는 현존하는 측정도구들, 적어도 사용 중인 것들이 부실하다고 결론 내렸습니다; 이들은 주로 그 질(신뢰도, 타당도, 공정성)에 대한 최소한의 기본 정보만 있거나 전혀 없는 등급척도 자기보고입니다. Stecher와 Hamilton(2014)은 지금까지 검토된 능력들을 "측정하기 어려운 역량"이라고 언급했습니다. 그들은 "학업적 마인드셋, 협력, 구두 의사소통, 학습하는 법 배우기, 그리고 다른 측정하기 어려운 21세기 능력과 역량과 관련된 명확하고 포괄적인 연구 의제를 개발할 필요가 있다"고 결론 내렸습니다(p. 71).

전통적인 검사는 아직 이러한 능력들을 일상적으로 측정하는 과제를 수행할 준비가 되어 있지 않지만, 고용주들(역사적으로)과 입학 담당자들(점점 더)은 이러한 능력들이 중요하다고 믿으며 따라서 면접, 이력서, 추천서, 또는 자기보고를 통해 이러한 능력들을 주관적으로 평가할 것입니다. 하지만, 지난 10년 동안 측정하기 어려운 능력을 측정하는 방법에서 상당한 발전이 있었습니다. 예를 들어, 협력과 협력적 문제해결을 측정하고 시뮬레이션을 사용하기 위한 컴퓨터 플랫폼이 설계되었습니다(Hao et al., 2024). 게임과 게임 기반 평가가 광범위한 능력(Landers & Sanchez, 2022)과 성격(Landers et al., 2022)을 평가하기 위해 개발되었고 운영 환경에서 점점 더 많이 사용되고 있습니다(Buckley et al., 2021). 상황판단검사는 이제 직업 현장에서 매우 일반적이며(OPM, n.d.) 교육 환경에서도 점점 더 증가하고 있고(Wolcott et al., 2020) 게임화될 수 있습니다(Landers et al., 2022). 우리는 이러한 발전을 혁신적 방법 섹션에서 검토합니다.

3.4 미래 능력 수요의 예측

직장에서 요구되는 능력의 특성이 기술로 인해 변화하고 있습니다—오늘날 가치 있는 많은 능력들이 곧 자동화될 것 같으며, 아직 인식되지 않은 새로운 능력들이 출현할 것입니다.

이러한 변화는 교육뿐만 아니라 노동력에도 영향을 미칠 것입니다. 하지만 어떤 능력이 단계적으로 사라지고 어떤 새로운 능력이 나타날 수 있는지 어떻게 알 수 있을까요? 누구도 미래를 확실하게 예측할 수는 없지만, 미래의 많은 부분이 현재와 비슷할 것이라고 가정하는 것이 안전합니다. 따라서, 미래의 학교와 직업 현장에서 요구될 능력이 대체로 우리가 방금 검토한 것들일 것이라고 가정하는 것이 유용한 출발점입니다. 그러나 여기서는 두 가지 추가적인 방법, 즉 직업 현장의 능력 요구사항에 대한 동향 분석과, 어떤 직업이나 직업의 일부가 기술에 의한 대체나 보완에 취약할 수 있는지를 결정하기 위한 직업의 과제 분석을 탐구할 것입니다.

3.4.1 동향 분석

여러 경제학 연구들이 노동시장에서 능력의 가치를 결정하기 위해 직장 동향을 조사했습니다. 미국 노동부의 직업 사전 데이터를 사용하여, Autor et al.(2003)은 1960년부터 1998년까지의 직업을 조사했고 그 기간 동안 기술이 일상적인 인지적 및 수동적 과제를 대체할 수 있었음을 발견했습니다. 기술은 또한 근로자들에게 새로운 인지적 요구를 부과했고 직장에서 가치 있는 것이 무엇인지에 영향을 미쳤습니다. 기술은 비일상적 작업과 대인관계 과제를 포함하는 활동을 보완하여, 수동적 작업과 일상적 인지 작업의 감소를 가져왔지만 다른 종류의 작업의 증가도 가져왔습니다.

비슷한 현상이 통신 기술(예: 인터넷, 소셜 미디어; 2000-2015)로 인해 발생하여 사회적 능력에 대한 새로운 요구를 부과했습니다(Deming, 2017). 사회적 능력은 더 효율적인 팀워크를 가능하게 합니다(Deming에 따르면, 근로자들은 비교 우위를 활용하기 위해 과제를 교환합니다). Deming(2017)은 2000-2012년 동안 가장 빠르게 성장한 직업이 교사, 관리자, 간호사, 치료사와 같은 사회적 직업이었음을 보여주었습니다. 엔지니어, 제도사와 측량사, 건축가, 생물학자와 물리학자와 같은 비사회적 STEM 직업은 부정적 성장을 경험했습니다. 사회적 직업은 미국의 모든 직업 중 비중이 12% 증가했으며, 임금도 더 빠르게 증가했습니다. Weinberger(2014)는 인지적 능력과 사회적 능력 중 하나만 요구하는 직업들에 비해 둘 다 높은 수준으로 요구하는 직업에서 고용과 소득의 성장을 발견했습니다.

Langer와 Wiederhold(2023)는 견습생들이 견습 기간 동안 받은 인지적, 사회적, 디지털, 수동적, 관리적, 행정적 능력 훈련의 수준을 나타내는 독일 견습 기록 데이터를 조사했습니다. 그들은 견습 1개월이 더 높은 임금과 관련하여 학교 교육 2-3개월의 가치가 있다는 것을 발견했습니다; 수익률은 디지털, 그 다음 사회적, 그 다음 인지적 능력 순으로 가장 높았습니다; 그리고 인지적 능력과 사회적 능력을 모두 증가시킨 견습이 가장 큰 수익을 제공했는데, 이는 이전 연구들(Deming, 2017; Deming & Kahn, 2018; Weinberger, 2014)과 일치하는 능력 보완성을 나타냅니다.

3.4.2 미래 직업에 대한 예측적 AI의 영향

AI가 미래의 직업에 미치는 영향에 대해 많은 글이 쓰여졌습니다. 편의상, 우리는 이러한 저작들을 두 단계로 나눌 것입니다: 첫 번째는 예측적 AI로도 불리는 기계학습에 초점을 맞춘 것이고, 두 번째는 대규모 언어 모델(LLM) 생성형 AI에 초점을 맞춘 것입니다.

AI의 파괴적 경제학에 대한 Agrawal et al.(2022)의 책은 예측적 AI의 가치라는 관점에서 쓰여졌습니다. 그들은 AI가 예측과 판단의 역할을 맡게 될 것이라고 주장했는데, 이는 Autor et al.(2003)이 관찰한 일상적 인지 과제의 자동화를 넘어 독특하게 인간의 능력을 필요로 한다고 가정되었던 복잡한 인지 과제로까지 상당히 확장됩니다. Agrawal et al.은 잠재적 시너지 이점과 그러한 이점을 달성하기 위해 업무가 어떻게 재조직될 수 있고 될 것 같은지에 초점을 맞췄습니다; 그러나 또 다른 함의는 이전의 기술 발전에서 노출된 저숙련 및 중숙련 직업이 아닌, 고숙련 직업조차도 AI에 노출될 것이라는 점이었습니다.

침투를 수치화한 최초의 연구 중 하나는 Frey와 Osborne(2017)이 수행했는데, 그들은 전문가 평가 연구를 통해 직업의 전산화 취약성을 연구했습니다. 전문가들은 ONET 데이터베이스에서 표본 추출한 70개 직업을 평가하여 어떤 직업이 완전히 자동화될 수 있는지 판단했습니다. 이를 바탕으로 그들은 컴퓨터나 로봇에 의한 자동화에 취약하지 않은 9개의 ONET 능력을 식별하고 이를 702개 직업의 더 큰 목록과 대조하여 노동력과 특정 직업의 자동화 취약성을 추정했으며, 고용의 47%가 위험에 처해 있다고 결론 내렸습니다. 9개의 비취약 능력은 타인 돕기와 돌보기, 설득, 협상, 사회적 통찰력, 순수 예술, 독창성, 수동 기민성, 손가락 기민성, 협소한 작업공간이었습니다.

더 최근의 전문가 평가 연구(Lassébie & Quintini, 2022)도 비슷하게 전문가 평가 접근법을 사용했지만, 전문가들은 직업 대신 능력과 역량의 자동화 가능성을 평가했는데, 이는 AI와 자동화가 직업에 미치는 영향을 더 정확하게 추정할 수 있게 했습니다. 자동화 가능성이 가장 낮은 O*NET 능력에는 인적 자원 관리, 복잡한 문제해결, 협상, 사회적 통찰력, 타인 돕기와 돌보기, 기술 설계, 물적 자원 관리, 적극적 학습, 서비스 지향성, 수리, 독창성, 설득, 적극적 경청이 있었습니다(반대쪽 끝에는 수리 능력, 암기, 손목-손가락 속도, 선택적 주의력, 정적 근력이 있었습니다). 결과는 대체로 Frey와 Osborn(2017) 연구와 일치했지만, Lassébie와 Quintini(2022)는 복잡한 문제해결과 적극적 경청을 추가하고, 순수 예술, 협소한 공간 작업, 손가락 기민성, 수동 기민성을 제외했습니다. 또한 여러 능력에 대해 의견 불일치가 있었는데, 여기에는 타인 상담과 조언(AI는 일부 맥락에서 이를 수행할 수 있음), 판매나 타인 영향(추천 시스템이 잘 수행함), 교수(일부 교수 활동은 AI가 잘 수행할 수 있음), 자신과 타인의 시간 관리(동적 일정관리 기술이 효과적이지만 적용 가능성이 제한될 수 있음), 구두 및 문서 표현(이 연구가 수행되었을 때도 자연어 처리[NLP]의 빠른 진전이 이 영역에서 이미 강력한 성능을 보여주고 있었음), 업무와 활동 일정관리(AI 과제 계획이 잘 발달됨), 시각적 능력(AI 비전은 지난 10년 동안, 특히 COVID 동안 상당히 발전함)이 포함됩니다.

OECD(2023)는 이 연구와 다른 연구들의 발견을 노동시장과 고용 전망의 더 큰 맥락에 놓기 위해 더 큰 연구를 수행했습니다. 그들은 AI가 노동시장에 상당한 영향을 미칠 것 같지만 그 영향이 무엇일지와 신뢰할 수 있는 사용을 촉진하기 위해 어떤 적절한 정책 조치가 필요할지에 대해 상당한 불확실성이 있다고 결론 내렸습니다. AI가 적절한 역량을 가진 고숙련 근로자를 위한 새로운 과제와 직업을 만들고 있으며 AI가 지루한 과제를 줄이고 참여와 안전을 증가시킬 수 있다는 징후가 있습니다. 하지만 이러한 변화는 또한 더 강도 높고 빠른 속도의 작업 환경을 남길 수 있습니다. AI 작업 관리는 인식된 공정성을 증가시킬 수 있지만 프라이버시를 위험에 빠뜨리고 편향을 도입하거나 영속화할 수 있습니다. 핵심적인 정책적 함의는 근로자들이 새로운 기술을 사용할 수 있는 능력을 갖추도록 보장하기 위한 교육과 훈련의 필요성이 증가하고 있다는 것입니다.

3.4.3 생성형 AI가 미래 직업에 미치는 영향

두 번째 단계는 2022년 11월 OpenAI의 ChatGPT와 2023년 3월 GPT-4의 출시 이후에 이어졌으며, 생성형 AI로 불리는 LLM에 초점을 맞췄습니다. Google의 Gemini, Anthropic의 Claude, Meta의 LLaMa를 포함한 다른 시스템들도 있습니다. Stable Diffusion, Midjourney, DALL-E를 포함한 텍스트-이미지 생성형 AI 시스템들도 있습니다. AI 전문가 커뮤니티는 상업적 출시 이전에 기반 기술의 발전을 일반적으로 인식하고 사용하고 있었지만(Lassébie & Quintini, 2022), Cotra(2023)가 지적했듯이 ChatGPT의 능력은 전문가 AI 커뮤니티조차 놀라게 했으며, 일부는 그 능력 수준에 대한 예측이 앞으로 10년이나 20년, 심지어 더 먼 미래에나 달성될 것이라고 제안했습니다.

ChatGPT와 관련 기술의 가능한 영향에 대한 한 연구는 OpenAI(Eloundou et al., 2023)에 의해 수행되었는데, 그들은 이것이 범용 기술(증기와 전기와 같은 범용 기술에 대한 검토는 Bresnahan, 2010 참조)이 될 잠재력이 있을 수 있다고 제안했습니다. 그들의 접근법은 노동 증강과 노동 대체 효과를 구분하지 않고 과제와 직업의 LLM에 대한 "노출"을 결정하는 것이었습니다. 그들은 주로 O*NET을 기반으로 과제의 LLM 노출을 측정하는 루브릭을 적용하기 위해 인간 평가자(주석자)와 GPT-4를 모두 사용했습니다. 그들은 현재 LLM 능력과 관련 도구를 고려할 때 직업의 19%가 그들의 과제의 최소 50%가 노출되어 있다고 결론 내렸습니다; 그러나 다른 생성 모델과 보완 기술을 추가하면, 근로자의 49%가 그들의 과제의 절반 이상이 노출될 수 있습니다. 능력과 관련하여, 과학과 비판적 사고에 의존하는 역할은 노출과 부정적 상관관계를 보이지만(즉, LLM에 덜 취약함) 프로그래밍과 글쓰기 능력은 LLM 노출과 긍정적으로 연관되어 있습니다(즉, LLM에 매우 취약함).

Eloundou et al.(2023)의 분석은 여기서 검토된 것과 같은 이전 연구들과 극적으로 다른 결론을 도출하지 않았습니다. Eloundou et al.은 그들의 노출 추정치와 다른 연구들에서 얻은 추정치 간의 상관관계를 계산했고 일반적으로 이들이 긍정적이고 통계적으로 유의미함을 발견했습니다. 그러나 그들의 분석에는 Lassébie와 Quintini(2022) 연구는 포함하지 않았습니다.

새로운 직업이나 새로운 능력의 출현에 대해서도 추측해 볼 수 있습니다. 예를 들어, 보조 기술(예: LLM)과 함께 일하는 것이 중요한 새로운 능력이 될 것 같습니다. 이미 디지털 문해력은 일부 고용주 설문과 다른 맥락에서 나타나는 능력입니다. 그러나 디지털 문해력은 "킨들로 읽기부터 웹사이트의 타당성을 평가하거나 YouTube 비디오를 만들고 공유하는 것까지 모든 것을 포함하는" 매우 광범위한 개념이어서(Loewus, 2016) 그 중요성에 대한 진술을 해석하기가 어렵습니다. 그러나 ChatGPT와 다른 생성형 AI 기술을 사용하는 것은 이미 가치 있는 능력이며, 기술이 발전함에 따라 생성형 AI 기술을 계속 사용하는 것이 중요한 능력으로 남을 것 같습니다. 개인 디지털 비서의 개념은 사라지지 않을 것 같습니다. 컴퓨팅 커뮤니티 컨소시엄과 인공지능발전협회의(Gil & Selman, 2019) 20년 로드맵은 개인 비서의 미래 세계를 구상합니다.

AI 프롬프트 엔지니어도 상당한 주목을 받았으며 새로운 능력 집합을 포함하는 새롭고 중요한 직업이자 미래의 최고 평가 직업으로 널리 선전되고 있습니다. 그러나 Acar(2023)는 이에 동의하지 않으며, 생성형 AI 시스템의 미래 버전이 더 직관적이 되고 신중한 프롬프트 작성에 덜 의존하게 될 것이라고 주장했습니다; 실제로, 그는 GPT-4와 같은 AI 모델이 프롬프트

엔지니어링에 매우 능숙하며 계속해서 더 나아질 것 같다고 주장했습니다. 또한, 프롬프트 엔지니어링은 LLM에 특화되어 있어 그 유용성이 제한적입니다. 대신, 그는 "문제를 식별, 분석, 서술하는 능력"이라는 의미의 문제 공식화가 LLM의 결과를 해석하고 비평하고 필요한 경우 재구성하여 다시 실행하는 능력과 함께 중요한 능력으로 등장할 것 같다고 제안했습니다. Acar는 문제 공식화의 네 가지 핵심 구성요소가 문제 진단, 분해, 재구성, 제약 설계이며 이러한 능력들이 AI 시스템과의 효과적인 협력에 핵심이 될 것이라고 제안했습니다.

3.5 결론: 미래를 위한 능력

지난 세기 동안 평가의 노력과 발전은 주로 교육과정 능력의 평가와 관련되어 왔습니다: 수학, 읽기, 과학, 즉 전통적인 K-12 교육과정이 목표로 하는 능력들입니다. 결과적으로, 이러한 능력들은 주와 국가의 교육 시스템을 모니터링하도록 설계된 대규모 국내 및 국제 평가의 초점이 되어왔습니다. 이러한 능력들은 중요하며 앞으로도 그럴 것이지만, 지난 20년 정도 동안 다른 종류의 능력들의 중요성에 대한 인식이 높아져 왔고, 이제는 최소한 동등하게 중요한 것으로 인정됩니다—협력, 문제해결, 비판적 사고, 창의성, 호기심, 직업윤리입니다. 때로는, 특히 직업 분야에서, 이들은 지속가능한 능력으로 불리는데, 이는 모든 종류의 교육, 훈련, 직무 및 맥락에서의 일반화 가능성과 유용성을 나타냅니다. 이러한 능력들은 측정하기가 더 어려워서 측정하기 어려운 능력으로 불릴 수 있습니다. 기술과 AI의 발전으로 어떤 능력이 가장 가치 있는지에 대한 변화가 계속될 것 같습니다. 이미 우리는 AI가 대학원생 이상의 수준에서 언어 과제, 예술 창작, 코딩 과제를 수행할 수 있음을 봅니다.

이러한 상황은 평가에 도전과 기회를 제시합니다. 도전과제는 오늘날 우리가 의존하는 단순한 자기평가 등급이 점점 더 중요해질 능력에 대한 유용한 정보를 제공하는 과제에 충분하지 않다는 것입니다. 기회는 우리가 오늘날 수학, 읽기, 과학을 측정할 수 있게 하는 것과 같은 수준의 정교함으로 측정하기 어려운 구인들을 평가할 수 있도록 새롭고 혁신적인 평가 방법을 개발할 수 있다는 것입니다.

4 혁신적 측정: 측정하기 어려운 기술을 평가하기 위한 새로운 접근

이 섹션에서는 측정하기 어려운 능력을 평가하는 현재의 방법을 검토합니다. 일부 능력에 대해서는 평가 도구가 개발되었지만, 많은 경우 여전히 **자기보고(Self-Report)와 타인 보고(Others' Reports)**에 의존하고 있습니다. 우리는 이러한 방법의 한계를 논의하고, 이를 개선할 수 있는 방법을 탐색합니다. 특히, 측정하기 어려운 능력을 평가하는 수행 평가(Performance Measures) 개발 노력에 초점을 맞춥니다. 여기에는 상황 판단 테스트(Situational Judgment Tests, SJTs), 게임(Game-Based Assessment), 시뮬레이션(Simulations), 상호작용 과제(Interactive Tasks) 등이 포함됩니다. 또한, 평가 대상자의 문제 해결 과정에서 발생하는 **행동 데이터(Process Data)**를 분석하여, 응답자의 행동 패턴, 반응 시간(Response Time), 대화 내용 등을 통해 능력을 추론하는 방법도 논의합니다.

4.1 논의의 기초 마련

이전 섹션에서는 미래에 점점 더 중요해질 가능성이 있는 기술(스킬)에 대해 논의했습니다. 이번 섹션에서는 이러한 기술을 측정하는 방법을 다룹니다. 기술과 측정 방법을 구분하는 것은 쉽지 않습니다. 예를 들어, 코딩과 같은 기술적 역량이나 수학, 독해력, 작문과 같은 전통적인 학문적 기술은 객관식 문제, 서술형 문제, 또는 기타 혁신적인 문제 형식을 통해 평가됩니다. 그러나 **대인관계 기술(소프트 스킬)**은 이러한 방식으로 평가하기 어려우며, 대신 **면접이나 평정 척도(Self/Other Reports)**와 같은 주관적인 방법에 의존하는 경우가 많습니다. 이러한 주관적 측정 방식은 정보의 신뢰성이 낮다고 여겨질 수 있습니다. LinkedIn Talent Solutions(2019) 보고서에 따르면, 인재 관리자의 91%가 소프트 스킬이 미래 채용에서 중요할 것이라고 생각했으며, 92%는 소프트 스킬이 하드 스킬만큼이나 중요하다고 응답했습니다. 하지만, 57%의 관리자들은 소프트 스킬을 정확하게 평가하는 것이 어렵다고 답했습니다. 현재 소프트 스킬 평가 방법으로는 행동 기반 질문(75%), 신체 언어 읽기(70%), 상황별 질문(58%) 등이 사용되고 있으며, 모두 주관적 요소가 강한 방식입니다.

표 8은 다양한 관점에서 평가 방법과 문항 유형을 정리한 목록을 제시합니다. Scalise와 Gifford(2006)는 Bennett(1993)의 연구를 기반으로, 컴퓨터 기반 평가의 문항 유형을 학업 과목 중심으로 분류한 체계를 제안했습니다. RAND(2020)의 평가 자료집은 K-12 교육 평가를 시험 유형별로 정리하여 교육 실무자가 쉽게 찾아볼 수 있도록 했으며, 사회·정서적 구성 요소까지 포함하도록 확장되었습니다. IMS Global(2022)의 QTI(Question Test Interoperability) 표준은 모든 디지털 평가를 지원하도록 설계되었으며, 상호작용(interaction) 유형을

포함하고 있습니다. 이는 본 섹션의 논의와 관련이 있기 때문에 목록에 포함되었습니다. 또한, Institute of Medicine(2015)과 미국 인사관리처(OPM, n.d.)의 평가 방법 목록은 각각 임상 평가 및 조직 채용과 선발에서 사용되는 방법을 제시하였으며, 해당 분야의 실무자들이 사용하는 용어를 반영하고 있습니다.

표 8에서 제시된 평가 방법 목록을 검토하면 몇 가지 문제가 명확해집니다. 평가에는 매우 다양한 접근 방식이 존재하며, 구성 개념(construct)과 측정 방법(method)이 종종 혼재되는 경우가 많습니다. 예를 들어, **Institute of Medicine(2015)**과 **미국 인사관리처(OPM, n.d.)**는 성격 검사(personality test)를 평가 방법으로 분류했는데, 이는 특정한 구성 개념이면서 동시에 평정 척도(rating scales)라는 방법론을 포함합니다. 마찬가지로, 인지 검사(cognitive tests)와 인지 능력(cognitive ability)도 평가 방법이면서 동시에 측정하려는 개념입니다. **RAND(2020)**의 목록은 주로 평가 유형(testing type)이라는 기준으로 정리되었지만, 평가할 내용(예: 대인(interpersonal), 개인(intrapersonal), 인지(cognitive))이라는 또 다른 차원이 존재합니다. Scalise & Gifford(2006) 및 IMS Global(2022)의 QTI 표준은 각각 다른 목적과 방식으로 구성 개념과 측정 방법을 분리하려는 노력을 했습니다. 이들은 개념과 상관없이 평가 응답을 어떻게 수집할 것인지에 초점을 맞추어 데이터 해석을 가능하게 하는 방법을 탐색했습니다. 이러한 **구성 개념-측정 방법의 분리(Construct-Method Separation)**는 **증거 중심 설계(Evidence-Centered Design, Mislevy et al., 2003)**의 핵심 요소 중 하나입니다.

원칙적으로, 동일한 기술(skill)을 여러 가지 방법으로 측정할 수 있습니다. 다특성-다방법 접근법(Multitrait-Multimethod Approach, Campbell & Fiske, 1959) 및 **모델링 프레임워크(Kyriazos, 2018)**는 이를 반영하기 위해 설계되었습니다. 예를 들어, **'호기심(Curiosity)'이라는 구성 개념(구인, Construct)**을 측정하는 다양한 방법이 존재합니다. - 자기보고(Self-Report): "나는 사물이 어떻게 작동하는지 아는 것을 좋아한다." (매우 그렇지 않다 ~ 매우 그렇다) - 교사 평가(Teacher Rating): "이 학생은 사물이 어떻게 작동하는지 알고 싶어한다." (참/거짓) - 컴퓨터 기록 분석(Computer Log Data): 학생이 새로운 옵션을 탐색한 횟수를 측정 - 수행 평가(Performance Test): 컴퓨터 게임에서 문을 연 횟수나 탐험한 경로 수 - 상황 판단 테스트(SJT): "5페이지짜리 연구 보고서를 작성해야 하는데, 조사 도중 관련 없는 흥미로운 연구 방법을 발견했다면 어떻게 하겠는가?" - 행동 기반 인터뷰(Behavioral Interview): "호기심이 발휘되어 흥미로운 것을 발견한 경험을 이야기해 주세요." 향후 평가 방식은 기술적·인지적 학문적 구인을 측정하기 위해 개발된 기존 평가 기법을 확장하여, 측정하기 어려운 기술(hard-to-measure skills)까지 포함하려는 시도가 증가할 것으로 예상됩니다.

Table 4.1: 다양한 관점에서 본 검사 방법과 문항 유형

인지적 문항 유형의 컴퓨터 기반 평가 (Scalise and Gifford [2006])	학업적, 사회적, 정서적 학습 (RAND [2020])	QTI 표준 (상호작용 유형) (IMS Global [2022])
객관식	지필	선택
선택/식별	디지털	텍스트 입력
재배열/재정렬	구두	확장형 텍스트
대체/교정	선택형 응답	빈칸 매칭

인지적 문항 유형의 컴퓨터 기반 평가 (Scalise and Gifford [2006])	학업적, 사회적, 정서적 학습 (RAND [2020])	QTI 표준 (상호작용 유형) (IMS Global [2022])
완성	자유 응답	핫스팟
구성	수행 과제	핫텍스트
발표		인라인 선택

심리학적 평가 측정과 방법 (Institute of Medicine [2015])	취업 적성 검사 (OPM [n.d.])	
선별 도구	성과 기록	매칭
체크리스트	평가 센터	그래픽 연결
설문지	이력서	미디어
기억력 검사	인지 능력	위치 객체
면접 관찰	정서 지능	선택점
관찰	청렴성/정직성 검사	슬라이더
인지 검사	직무 지식 검사	업로드
등급 척도	성격 검사	그리기
	신원 조회	사용자 정의
	상황 판단 검사	종료 시도
	구조화된 면접	
	훈련과 경험	
	작업 샘플	

Roll & Barhak-Rabinowitz(2023)은 측정하기 어려운 능력 중 하나인 **자기 조절 학습(Self-Regulated Learning, SRL)**을 PISA 2025 디지털 환경에서의 학습(Learning in a Digital World, LDW) 평가에서 측정하는 방안을 제안했습니다. SRL은 인지 및 메타인지 과정, 정서 조절, 동기를 포함하는 복합적인 개념입니다. 현재 SRL을 측정하는 가장 일반적인 방법은 **자기보고 설문(Self-Report Questionnaire)**이지만, 이는 개인이 평가 기준을 다르게 적용하면서 발생하는 **기준 편향(Reference Bias)**의 문제가 있습니다. 이를 해결하기 위해, Roll & Barhak-Rabinowitz는 SRL을 측정할 새로운 틀을 제안했습니다. 이 접근법은 학습자가 학습 활동 중 실제로 수행하는 행동을 기반으로 평가하는 것입니다. - 실험하기(Experimenting): 상호작용이 가능한 시뮬레이션을 통해 학습자가 직접 탐색 - 피드백 받기(Receiving Feedback): 자동으로 피드백을 받거나 버튼을 눌러 요청 가능 - 정보 탐색(Seeking Information): 튜토리얼 시청, 힌트 요청, 예제 보기 등을 통해 학습 이러한 요소들을 PISA LDW 평가에서 어떻게 적용할 수 있을지 체계적으로 정리했습니다.

Roll & Barhak-Rabinowitz(2023)의 연구는 복잡한 기술을 측정하기 위한 혁신적 평가(Innovative Assessments for Complex Skills) 관련 논문 모음집(Foster & Piacentini, 2023)의 일부입니다. 이 연구 모음의 핵심 요점은 다음과 같습니다. - 쉬운 것이 아니라

중요한 것을 측정해야 한다 → 기존의 단순한 시험보다 실질적으로 의미 있는 평가가 필요함. - 평가는 실제 맥락에서 이루어져야 하며 학습과 연결되어야 한다 → 평가가 실생활과 유사해야 함. - 평가 설계의 모든 과정에서 혁신이 필요하다 → 문항 개발, 채점 방식, 결과 해석 등 전반적인 혁신 필요. - 디지털 기술을 활용하면 더 많은 것을 측정할 수 있지만, 더 정교한 측정 모델이 필요하다 → 기술 발전에 맞춰 평가 방식도 개선해야 함. - 평가의 타당성이 중요하다 → 평가 결과가 신뢰할 수 있도록 검증 과정이 필요함.

이러한 주제들은 모두 미래 평가 연구에서 중요한 우선순위를 차지하지만, 각각의 적용 분야에 따라 연구의 초점이 다를 수 있습니다. 향후 연구에서는 중요하지만 측정하기 어려운 기술(hard-to-measure skills)에 대한 관심이 더욱 증가할 것으로 예상됩니다. 실제 학습 맥락(authentic learning context)에서의 평가 개념은 새로운 것이 아니지만(Erwin & Sebrell, 2003; Frensch & Funke, 1995), 기술 발전이 이러한 평가 방식의 유용성을 더욱 높일 가능성이 있습니다. 또한, 학습을 평가하는 개념도 새로운 것은 아닙니다. **역동적 평가(dynamic assessment)**에 대한 연구는 오랜 역사를 가지고 있으며(Grigorenko & Sternberg, 1998), 심리학(Bolsinova et al., 2022; Deonovic et al., 2018; Yeung, 2019)과 경제학(Heckman & Zhou, 2021)에서도 유망한 새로운 측정 접근법이 지속적으로 연구되고 있습니다.

이 섹션에서는 **측정하기 어려운 능력(hard-to-measure skills)**을 평가하는 주요 방법을 검토합니다. 평가 방식은 다음 네 가지로 구분됩니다. - 평정 및 순위 평가(Ratings & Rankings) - 상황 판단 테스트(Situational Judgment Tests, SJTs) - 수행 평가(Performance Measures) - 다중 양식 평가(Multimodal Measures)

4.2 평정 및 관련 방법

평정 방식은 평가자가 자신이나 타인의 특성을 평가하는 방법으로, 일반적으로 리커트 척도와 같은 평정 척도를 사용하지만, 체크리스트와 같은 변형된 형태도 존재한다. 특히 자기 평정 방식은 매우 널리 활용되며, 심리적·교육적 개념 전반에 걸쳐 적용할 수 있을 만큼 유연하고, 개발·시행·채점·결과 보고가 비교적 저렴하기 때문에 높은 인기를 끌고 있다. 평정 척도 방식에 대한 심리 측정 모델과 개념적 네트워크도 이미 잘 구축되어 있으며, 향후에도 다양한 기술과 역량을 측정하는 데 계속해서 중요한 역할을 할 것으로 예상된다. 실제로 성격(John & Srivastava, 1999), 흥미(Su et al., 2019) 등과 같은 심리학적 개념의 많은 부분이 평정 척도를 기반으로 연구되어 왔다.

그러나 평정 척도 방식에는 몇 가지 한계가 존재한다. 자기 보고 방식은 응답 스타일 편향(van de Vijver & He, 2016), 참조 기준 편향(Lira et al., 2022), 사회적 바람직성 편향(Paulhus, 2002), 그리고 의도적인 조작(Geiger et al., 2021)과 같은 문제에 취약하다. 특히, 입학 전형이나 채용 평가와 같은 고위험(high-stakes) 상황에서는 이러한 문제가 더욱 심각하게 나타날 수 있다(Niessen et al., 2017).

타인 평가 방식(정보 제공자 평정)은 사회적 바람직성 응답과 조작 가능성을 어느 정도 완화할 수 있다. 물론 평가자가 평가 대상에게 유리하도록 과장된 응답을 할 가능성도 있지만, 전반적으로 자기 보고 방식보다 미래 행동을 더 잘 예측하는 경향이 있다(Connelly & Ones, 2010; Oh et al., 2011; Poropat, 2014). 예를 들어, 추천서는 학업 성취도와의

직접적인 상관관계는 낮지만, 학위 취득과 같은 장기적 성과를 예측하는 데는 유용한 것으로 나타났다(Kuncel et al., 2014).

이러한 한계를 보완하기 위해 순위 방식과 앵커링 기법이 대안으로 제시되고 있다.

순위 방식 중 하나인 강제 선택(Forced-choice) 방식은 응답자가 특정 항목을 단순히 평정하는 것이 아니라, 항목 간 우선순위를 정하도록 요구함으로써 사회적 바람직성 응답과 응답 스타일 편향을 줄이는 데 효과적이다. 특히 입학 전형과 같은 고위험 평가에서 유용하며, 최근 강제 선택 방식의 채점 방법이 발전하면서 신뢰도가 더욱 향상되었다(Fu et al., 2024). 또한, 기존의 평정 척도 방식보다 더 높은 예측력을 보이는 것으로 나타났다(Cao et al., 2015; Salgado & Tauriz, 2014).

또 다른 접근법인 앵커링 기법은 응답 스타일 편향을 줄이는 데 초점을 맞춘다. 예를 들어, 앵커링 비네트(Anchoring vignettes) 기법은 응답자가 자신뿐만 아니라 가상의 인물도 함께 평가하도록 함으로써, 응답을 보다 객관적으로 조정할 수 있도록 한다(King & Wand, 2007). 이러한 방식은 국가 간 응답 스타일 차이를 줄여 개념의 비교 가능성을 높이는 데 효과적인 것으로 나타났다(Kyllonen & Bertling, 2013). 한편, Ludlow et al. (2022)는 유사한 기법을 활용하여 측정하기 어려운 역량인 '삶의 목적'을 평가하는 방법을 제시하였다. 타인 평가 방식에서는 행동 기반 평정 척도(BARS) 또한 널리 활용된다. BARS는 평가 기준을 명확히 제시하여 평가자가 일관된 판단을 내릴 수 있도록 돕는 방식으로, 주로 조직 평가에서 사용된다(Kell et al., 2017; Klieger et al., 2018). 하지만 BARS는 일반적으로 자기 평가보다는 타인 평가에 사용되는 경향이 있다.

결론적으로, 평정 척도 방식은 여전히 중요한 측정 도구이지만, 여러 가지 한계를 가지고 있으며, 이를 보완하기 위한 다양한 대안적 접근법이 개발되고 있다. 향후에는 보다 정교한 평가 기법이 도입됨으로써 측정의 신뢰도와 타당성이 더욱 향상될 것으로 기대된다.

4.3 상황판단검사

상황 판단 검사(SJT)는 상황 설명을 제시하고 응답자에게 해당 상황에 어떻게 대처할 것인지 또는 최선의 대응은 무엇인지 묻습니다. 그림 4는 그 예시를 제공합니다. SJT는 특히 대인 관계 기술과 같이 측정하기 어려운 구성 개념을 측정하는 데 널리 사용되는 방법입니다 (Christian et al., 2010). SJT는 유연한 방법으로 서면 자료나 비디오를 포함할 수 있으며, 일반적으로 응답 옵션의 순위를 매기거나 평점을 매기도록 요청합니다. SJT는 직원 선별 및 때로는 교육을 위해 조직 환경에서 널리 사용됩니다 (OPM, n.d.; Cox et al., 2017).

상황 판단 검사(SJT)는 교육 환경에서도 활용되어 왔습니다 (MacCann & Roberts, 2008; Sternberg et al., 2000). College Board는 학부 입학 시험으로 SJT를 실험했으며 (Schmitt et al., 2009), SJT는 경영 대학원 (Hedlund et al., 2006) 및 치과 대학원 입학 (Buyse & Lievens, 2011)에도 사용되었습니다. 미국 의과대학 협회(AAMC)는 현재 대인 관계 기술, 문화적 인식, 문화적 겸손, 공감 및 연민, 팀워크 및 협력, 자신과 타인에 대한 윤리적 책임, 회복력 및 적응력, 신뢰성 및 의존성, 학습 및 성장에 대한 헌신을 포함한 9가지 전문 역량을 측정하기 위해 의과대학 입학을 위한 75분짜리 SJT인 AAMC PREview Professional Readiness Exam을 제공합니다

(AAMC, n.d.). Acuity Insights (n.d.)는 14개의 시나리오(8개의 서면, 6개의 비디오)를 통해 사회 지능 및 전문성의 10가지 측면을 측정하는 90분짜리 개방형 SJT인 Casper라는 경쟁 시험을 제공하며, 매뉴얼을 발행합니다 (Acuity Insights, 2023).

상황 판단 검사(SJT)는 활용의 유연성과 측정하기 어려운 기술을 평가하는 데 적합하다는 점에서 그 가치를 입증했으며, 따라서 미래에도 평가 방법으로 인기를 유지할 가능성이 높습니다. 그러나 과제도 있습니다. SJT는 단위 시간당 검사 시간 대비 평점 척도 측정보다 신뢰성이 낮은 경향이 있으며, 신뢰할 수 있는 점수를 얻기까지 더 오래 걸리거나 더 많은 검사 시간이 필요합니다. 예를 들어, Casper는 90분짜리 시험이지만 사회 지능 및 전문성의 10가지 측면이 아닌 단일 요인만 측정합니다. 이는 SJT의 일반적인 특징입니다. 예를 들어, Oswald et al. (2004)은 12가지 고등 교육 역량(예: 리더십, 예술적 재능)을 개발했지만, 그들이 개발한 SJT는 단일 차원만 측정했습니다 (Schmitt et al., 2009). SJT 연구 및 홍보 자료에서 SJT로 여러 차원을 측정하려는 욕구를 나타내는 점을 고려할 때, SJT의 일반적인 미래 연구 과제는 합리적인 시간 내에 여러 차원을 안정적으로 측정하는 것이 될 것입니다.

You're one of the managers for a large volunteer agency. In a discussion about how to find new volunteers, you bring up what you think is a great new idea. But the other managers tell you that the idea is "off base" and not workable. How would you handle this situation?

A. Drop your idea because the group is probably right.
B. Point out several good reasons why your idea might work.
C. Drop your idea for now but tell it to your boss later.
D. Tell the other managers that lots of people don't recognize great ideas at first.

Figure 4.1: 상황판단검사(SJT) 문항 예시. 출처: Zu와 Kyllonen (2020)

4.4 수행 측정

미래를 위한 기술 섹션에서 확인된 주요 기술 중 일부에 대한 수행 평가는 비판적 사고(Liu et al., 2016) 및 창의성(Weiss et al., 2021)과 같은 경우 비교적 잘 확립되어 있습니다. 미래의 평가는 이러한 측정 방법을 응용 분야에 포함하고 추가적인 점진적 발전을 볼 수 있을 것입니다. 이러한 기술 측정의 또 다른 중요한 발전은 이러한 기술이 시간이 지남에 따라 어떻게 증가하는지와 관련될 것입니다. Koedinger et al. (2023)은 지능형 튜터링 시스템과 함께 배포된 수학, 과학 및 언어 수업의 100만 건 이상의 관찰 데이터에 기반하여 초기 지식 수준을 고려했을 때 학습 속도에서 규칙성을 발견할 수 있었습니다. Duolingo의 Birdbrain 시스템은 문항 반응 이론을 사용하여 능력과 문항 난이도에 따라 학생의 수행 능력을 예측하고, 적응형 테스트에서 해당 과정이 수행되는 방식과 유사하게 연습 수행 후 학생 능력 수준을 업데이트함으로써 참여와 학습을 위한 최적의 난이도 수준으로 어학 교육을 조정합니다 (Bicknell et al., 2023). 적응형 테스트 및 적응형 교육과 하이브리드 접근 방식에서와 같이 기술의 업데이트와 관련된 개념과 전통적인 기술 평가의 병합은 전자 교육 및 기록 유지가 더욱 보편화됨에 따라 미래 평가에서 점점 더 중요해질 가능성이 높으며, 특히 적응형 교육 응용 분야에서 하이브리드 문항 반응 이론 모델링의 사용도 중요해질 것입니다 (Scalise et al., 2023; Yeung, 2019).

또 다른 중요한 추진력은 현재 주로 체크리스트와 평점 척도로 측정되는 구성 개념 및 기술, 예를 들어 질문과 평점 루브릭을 통해 자기 동기, 독창성 및 시간 관리와 같은 다양한 지원자의 자질을 평가하는 일반적인 채용 면접에 대한 수행 평가 개발에서 나올 것입니다. 우리가 수행 측정에 적합하다고 생각하는 구성 개념의 종류에는 팀워크, 협업, 리더십, 자기 관리 및 자기 조절, 감정 관리, 직업 윤리, 유연성, 문화적 감수성 및 표 8에 나열된 기타 소프트 스킬 또는 지속 가능한 기술이 포함됩니다. 수행 측정을 통해 이러한 소프트 스킬을 측정하려는 시도는 오랜 전통을 가지고 있으며, 때로는 객관적인 성격 검사(Cattell & Warburton, 1967; Ortner et al., 2006)라는 이름으로 불리기도 합니다. Alan et al. (2019)의 그릿 게임과 내적 동기 측정으로서 코딩 속도 테스트에서 끈기를 사용한 Segal (2012)의 연구는 성격 특성의 수행 측정의 예입니다. Charness et al. (2018)는 끈기, 자기 관리 또는 성실성의 특성 측정으로 이해될 수 있는 행동 경제학 연구에서 사용된 실제 노력 과제 목록을 제공했습니다. Kyllonen과 Kell (2018)은 이러한 문헌의 대부분을 요약하여 저위험 인지 테스트(Segal, 2012), 객관적인 성격 테스트(Ortner & Proyer, 2015), 경제적 선호도 과제(Falk et al., 2018), 확신 판단(Stankov et al., 2015), 설문 조사 행동(Soland & Kuhfeld, 2021), 문항 위치 효과(Weirich et al., 2017) 및 반응 시간에서 추론된 노력(Wise, 2017)의 범주로 나누었습니다. 이 모든 것은 평점이 아닌 수행 과제를 통해 소프트 스킬을 측정하려는 시도로 이해될 수 있습니다.

협력적 문제 해결은 의사 소통, 팀워크 및 협업과 같은 소프트 스킬 또는 소프트 스킬 세트의 수행 측정의 한 예입니다. ETS는 협력적 평가 및 학습을 위한 ETS 플랫폼(EPCAL; Hao et al., 2017)과 협상(Martin-Raugh et al., 2020), 문자-숫자 문제 해결, 숨겨진 프로필 의사 결정(Kyllonen et al., 2021) 등을 포함한 일련의 과제를 개발했으며, 이는 팀 성과와 개인의 협업 기술을 모두 측정합니다(Hao et al., 2019). 미래 인력에서 사회적 기술의 중요성을 고려할 때(Deming, 2017), 이러한 종류의 평가 프로젝트는 미래에 중요성이 커질 가능성이 높습니다.

4.5 생활 자료(L-데이터)

Cattell (1965)은 "실제, 일상생활에서의 행동"으로 정의되는 L-데이터(생활 데이터)를 대상의 기술에 대한 추론의 근거로 사용하는 것을 제안했습니다. 행정 기록, 소셜 미디어, 휴대폰, 웹사이트 등 모든 종류의 기록을 광범위하게 이용할 수 있게 되면서, Cattell이 L-데이터 사용을 제안했을 때보다 L-데이터 수집이 훨씬 쉬워졌습니다. 이러한 종류의 데이터는 설문지나 SJT 응답에서 얻은 데이터와는 상당히 다릅니다. 이 데이터는 개인의 속성에 대한 추론을 이끌어낼 수 있는 개인들이 남긴 흔적과 지표입니다. 연구에서는 행정 기록을 사용하여 학생들의 학업 및 비학업적 기술과 행동을 반영하는 종합 점수를 생성하여 졸업을 예측하고 교사의 효과를 평가하는 데 사용했습니다 (Jackson, 2018; Kautz & Zanoni, 2014; Novarese & Di Giovinazzo, 2013). 소셜 미디어 및 기타 행동 흔적은 성격의 반영으로 간주되었습니다 (Gosling et al., 2011; Kosinski et al., 2014; Youyou et al., 2015). 예를 들어, Gosling et al. (2002)은 기숙사 방의 외관과 내용물을 기반으로 성격을 측정했습니다. 배경과 경험은 전통적인 설문 조사, 구조화된 이력서 또는 바이오 데이터(Mumford & Owens, 1987), 외래 평가(Trull & Ebner-Priemer, 2013) 및 휴대폰, 웨어러블 및 비콘의 모바일 감지 데이터(Mattingly et al., 2019; Mirjafari et al., 2019)를 사용하여 측정할 수도 있습니다. 미래의

평가는 학생과 근로자의 지식, 기술, 능력, 행동, 가치 및 태도에 대한 추론을 도출하는 데 사용될 수 있는 다양한 종류의 데이터를 통합하는 것을 점점 더 포함할 수 있습니다. 그러나 2024년 유럽 연합 AI 법, 미국 및 기타 지역에서 예상되는 규제, 그리고 많은 기관에서 발표하는 일반적인 AI 윤리 성명(Abrams, 2024; Blackman & Ammanath, 2022)을 고려할 때 개인 정보 보호 문제는 해결해야 할 것입니다.

4.6 게임 기반 접근법

게임 기반 평가는 "[응시자]가 핵심 게임 플레이 루프에 참여하는 플레이어인 동안 특성 정보가 추론되는 평가 방법"으로 정의할 수 있습니다 (Landers & Sanchez, 2022, p. 1). Landers와 Sanchez (2022)는 또한 게임화된 평가를 게임 메커니즘 또는 게임 개념이 기존의 전통적인 평가에 적용된 평가로, 게임 방식으로 설계된 평가를 테스트 개발자가 새로운 테스트를 설계할 때 게임 개념을 사용한 평가로 정의했습니다. 게임 기반 또는 게임화된 평가가 사용될 수 있는 몇 가지 이유가 있습니다. Landers와 Sanchez가 초점을 맞춘 응용 분야인 직원 선발과 같은 중요한 목적을 위해 평가는 지원자의 기술을 측정하는 데 도움이 될 뿐만 아니라 잠재적으로 현실적인 직무 미리보기 또는 후보자에게 "매력적인 경험"을 제공하는 채용 수단으로 사용될 수 있습니다 (Landers & Sanchez, 2022, p. 21). 게임 기반 평가는 호기심 및 사회적 선호도(Tang & Kirman, 2023)와 같이 측정하기 어려운 특정 기술을 고유하게 측정할 수도 있습니다. 세 번째 이유는 시험 응시자의 참여를 높이는 것입니다. 중요한 시험 상황의 응시자는 이미 시험 응시에 참여할 충분한 동기를 가지고 있지만, 학교 책임성 시험, 국내외 대규모 평가 및 연구 환경과 같이 응시자에게 중요도가 낮은 설정에서는 그렇지 않습니다. 이러한 설정에서 동기 부족은 점수에 영향을 미칠 수 있으며(Liu et al., 2012), 게임화된 버전의 테스트는 참여와 동기를 증가시켜 결과적으로 응시자의 기술을 더 잘 반영할 수 있습니다. Buckley et al. (2021)은 SimCityEDU: Pollution Challenge (Mislevy et al., 2014), ACTNext의 Crisis in Space (Chopade et al., 2019) 및 Imbellus의 Project Education Ecosystem Placement (PEEP)을 포함한 교육 분야의 게임 기반 및 게임화된 평가를 검토했습니다.

4.7 다중양식 측정 또는 과정 데이터

다중 모드 측정은 생리적 데이터(예: EEG, 심박수), 로그 파일에 기록된 행동 데이터(예: 대화, 채팅, 키 입력, 시선 추적) 및 인간 평가자가 분석하거나 자동으로 분석하는 오디오 및 비디오 녹화에서 포착된 자세와 표정을 사용하는 것으로 정의할 수 있습니다 (Molenaar et al., 2023; Slavich, 2019). 이러한 데이터는 평가 응용 분야에서 사용되기 시작했으며 그 사용은 증가할 가능성이 높습니다 (Martin-Raugh et al., 2023). 예를 들어, Chen et al. (2014)의 ETS 프로젝트는 17명의 발표자에게 4가지 발표 과제를 부여하고 오디오, 비디오 및 3D 장치를 사용하여 수행 능력을 포착하여 대중 연설 기술을 분석했습니다. 그들은 NLP 방법, 음성 처리 및 (Microsoft의 Kinect[Zhang, 2012] 사용) 음성 및 비언어적 의사 소통(손짓 및 머리 방향 포함)을 포착하는 다중 모드 감지를 사용하여 특징을 추출했습니다. Chen et al. (2014)는

추출된 데이터에 대한 채점 모델을 개발하고, 결과 점수가 대중 연설 수행에 대한 인간의 전체적인 평가와 상관관계가 있음을 발견했습니다.

다른 두 가지 ETS 프로젝트에서 Martin-Raugh et al. (2020)과 Jiang et al. (2023)은 협상 및 협력적 문제 해결 과정에서 대화를 분석하여 성공적인 협업과 덜 성공적인 협업 간의 처리 차이에 대한 통찰력을 얻었습니다. 그들은 자연어 처리 방법을 사용하여 대화 내용을 인사, 정보 공유, 기여 인정 및 협상과 같은 범주로 구성된 루브릭으로 분류하고, 개인과 팀이 수행한 대화의 성격과 과제 성공 간의 상관관계를 발견했습니다. 평가의 역사는 대부분 제한된 상호작용과 시험과 응시자 간의 매우 제한된 데이터 스트림에 기반해 왔습니다. 다중 모드 평가는 시험 응시자가 알고 있고 할 수 있는 것에 대한 추론을 도출할 수 있는 훨씬 더 풍부한 형태의 표현에 대한 문을 엽니다.

4.8 결론: 혁신적 측정

미래를 위한 기술 섹션에서 검토된 종류의 기술을 측정하는 주요 방법은 평점 척도법입니다. 해당 분야는 강조된 기술이 측정하기 어려운 기술이기 때문에 이 방법을 채택했으며, 평점 척도는 정의를 명확하게 설명할 수 있는 거의 모든 기술을 측정하기 위한 일반적이고 유연한 접근 방식입니다. 자기 보고식 평점을 개선하는 방법이 있습니다. 타인 보고는 준거 편향(Lira et al., 2022)과 같은 자기 보고와 관련된 편향에 덜 민감하지만, 후광 효과(Cooper, 1981)와 같은 자체적인 제한 사항이 있습니다. 강제 선택 측정 또한 자기 보고와 관련된 편향을 줄이므로 일반적으로 자기 보고보다 선호됩니다. SJT는 많은 측정하기 어려운 기술에 적용할 수 있는 또 다른 유연한 측정 방법입니다. 이는 일반적인 행동 평가가 아닌 정확하거나 적절하거나 유용한 행동에 대한 지식을 측정할 수 있다는 점에서 자기 보고보다 개념적 이점을 갖습니다. 미래의 평가는 자기 보고에서 이러한 다른 형태의 측정으로의 보다 일상적인 전환을 포함할 가능성이 높습니다.

그러나 우리는 미래 평가의 더 중요한 움직임은 미래의 중요한 기술을 측정하기 위해 게임과 같은 수행 기반 측정과 실제 협상 세션 또는 협력적 문제 해결 과제와 같은 상호 작용적 과제의 개발 및 채택을 포함할 것이라고 믿습니다. 성격의 수행 측정은 해당 분야에서 오랫동안 추구해 온 목표였으며(Ortner & Proyer, 2015), 일부 진전이 이루어졌습니다(Kyllonen & Kell, 2018; Linzarini & Catarino da Silva, 2024). 수행 측정은 원칙적으로 평점에 비해 상당한 이점을 갖습니다. 수행 측정은 후광, 준거 편향, 반응 양식, 사회적 바람직성과 같은 평점 편향에 영향을 받지 않으며, 행동에 대한 주관적인 평가보다는 객관적인 행동 표본이 될 수 있습니다. (그러나 관찰자 평점을 요구하는 수행 과제는 여전히 후광/뿔 효과[Noor et al., 2023], 엄격/관대함[Cheng et al., 2017] 및 편류[McLaughlin et al., 2009]와 같은 평점 편향의 영향을 받을 수 있습니다.) 수행 측정이 아직 많은 중요한 구성 개념에 대해 잘 개발되지 않았기 때문에 이러한 구성 개념은 측정을 위해 주관적인 평점에 계속 의존합니다. 우리는 수행 측정이 사용자, 학생 또는 직원의 기술 수준에 대한 추론을 도출하는 데 사용될 수 있는 프로세스 분석 및 데이터 마이닝을 포함하는 시험 없는 측정에 의해 보완될 것이라고 믿습니다(Baker & Yacef, 2009). 사회적 및 정서적 학습(Jackson, 2018; Kautz & Zanoni, 2014)에서 학업 성취도(Waheed et al., 2020) 및 STEM 직업 참여(Yeung & Yeung, 2019)에 이르기까지 다양한 영역에서 좋은 사례가 있습니다.

5 운영 혁신: 인공지능과 기술을 통한 도약

이 섹션에서는 시험 개발 주기의 단계를 다룹니다. 시험 목적의 초기 구상, 시행 및 행정적 제약 조건 고려부터 문항 개발, 시험 구성, 보안, 품질 관리, 채점, 그리고 타당성 및 공정성 고려를 포함한 시험 평가에 이르기까지 전반적인 과정을 살펴봅니다. 이 섹션의 주요 주제는 기술 발전, 특히 인공지능(AI)의 발전이 시험 개발의 모든 운영과 단계에 상당한 영향을 미칠 가능성이 크다는 점입니다.

5.1 논의의 기초 마련

시험 운영은 (미래를 위한 기술 섹션에서 정의된) 구성 개념과 (혁신적인 측정 섹션에서 정의된) 시험 방법을 고려하여 시험 양식을 시행하는 것과 관련된 시험 개발의 모든 단계를 의미합니다. 이러한 단계에는 시험 목적 및 행정적 제약 조건에 맞춘 설계, 문항 개발, 시험 양식 구성, 시험 검토, 시험 전달 및 시행, 채점, 점수 보고, 시험 평가, 문항 뱅킹, 시행에서 채점에 이르기까지 보안의 모든 측면 관리, 그리고 전체 과정에 대한 품질 관리 수행이 포함됩니다. 이러한 과정은 표준화된 시험 산업의 핵심입니다. Schmeiser와 Welch (2006)는 이러한 단계가 전통적으로 어떻게 수행되었는지와 주요 문제점을 포괄적으로 개관했으며, International Test Commission (2001, 2013, 2017)과 Association of Test Publishers (2022)는 기술 기반 평가(TBA)와 관련된 최근 고려 사항을 추가하여 이 연구를 보완했습니다. Schmeiser와 Welch는 지난 60년 동안 시험 개발이 예술에서 과학으로 진화하는 과정을 설명했습니다. 이러한 진화는 예를 들어 시험 구성의 비공식적인 경험 법칙에서 최근의 혼합 정수 프로그래밍 사용(Davey, 2023; van der Linden, 2005)으로, 또는 인간 채점에서 기계 채점으로 에세이 채점 방식이 변화한 것(Shermis & Burstein, 2013) 등 다양한 단계에서 확인할 수 있습니다.

우리는 미래의 평가가 기술, 운영 연구 방법, 그리고 예측 및 생성적 AI 방법 모두의 발전을 활용하여 시험 개발의 예술에서 과학으로의 전환을 지속하고 아마도 가속화할 것이라고 믿습니다.

이 섹션에서는 상당한 혁신을 볼 수 있는 기반을 제공하기 위해 시험 개발 단계의 현재 최첨단 발전을 검토합니다. 운영 영역 내의 일부 유망한 아이디어는 기술이 평가의 공정성과 평가를 통해 얻을 수 있는 정보의 신뢰성 모두에서 발전을 가능하게 할 것이라는 기대를 갖게 합니다. 공정성 처리는 전통적으로 시험 데이터가 수집된 후 항목 반응이 다른 시험 응시자 그룹에 대해 동일하게 해석될 수 있는지 통계적으로 평가하기 위해 평가 개발의 후반부에 가장 많이 개발되었습니다 (Millsap, 2011). 기술은 과거에는 비공식적인 정책과 체크리스트에 의존했던 항목 생성의 초기 단계로 공정성 정교함을 가져올 것을 약속합니다 (ETS, 2014, 2022). 보안

또한 기술이 평가 점수가 알 수 없는 영향에 의해 손상되지 않고 의도된 목표 구성 개념에 대한 시험 응시자의 기술을 직접적으로 나타내도록 보장하는 데 이점을 가져올 수 있는 영역을 나타냅니다. 이 섹션에서 이러한 문제를 다룹니다.

5.2 검사 실시와 행정적 제약 조건건

5.2.1 시간

평가의 일반적인 원칙은 시험에 더 많은 시간을 할애하거나 개인으로부터 더 많은 정보를 얻을수록 평가의 신뢰도가 높아진다는 것입니다. 그리고 정보가 더 신뢰할 수 있을수록 측정에서 도출된 추론이 정당화되고 유용할 가능성이 높아집니다. 예를 들어 미래 결과를 예측하는 데 유용합니다. 추가 정보가 있으면 측정 오차 잡음 속에서 구성 개념 신호가 점점 더 명확하게 나타납니다. 이러한 결과는 다른 모든 조건이 동일할 때, 사격 연습에서 체중 측정에 이르기까지 모든 평가 영역에서 참입니다. 즉, 시험이 문항 수나 시간 면에서 길수록 좋습니다. 대부분의 경우 문제는 개인들이 긴 시험을 치르기를 원하지 않고, 후원 기관이 긴 시험을 시행하거나 비용을 지불하기를 원하지 않는다는 것입니다.

이 문제를 해결하기 위한 몇 가지 전략이 있습니다. 시험 문항이 제공하는 정보량뿐만 아니라 시간도 고려하여 전통적인 심리 측정 접근 방식을 통해 시험을 보다 효율적으로 만들 수 있습니다. 이는 시험 시간의 50% 절감을 약속했던 적응형 검사의 주요 동기 부여 요인이었습니다 (van der Linden & Glas, 2010). 관련 시험(또는 척도)의 수행 능력을 사용하여 현재 응시 중인 시험(또는 척도)의 점수 추정치를 업데이트하는 다차원 적응형 검사(Segall, 1996)는 이 아이디어를 한 단계 더 발전시켜 동일한 측정 정밀도에 대해 추가적으로 33%의 시간 절감을 약속합니다. 이 아이디어의 미래 적용은 개인에 대한 모든 가용 정보 소스(소셜 미디어, 교육 기록, 추천서, 이력서, 자발적으로 제출된 자료)를 개인 정보 보호 제한 하에 단순히 기술 추정의 시작점으로 활용하여 시험 또는 평가 세션의 새로운 정보로 업데이트하는 것입니다. 이는 시험 세션에서 추가적인 시간 절약을 가져올 수 있습니다.

또 다른 전략은 시험 응시 경험을 응시자에게 더 유용한 방식으로, 즉 반환되는 이점을 통해 만드는 것입니다. 예를 들어, 교육은 일반적으로 평가와 함께 이루어지지만, 교육을 받는 학습자는 교육 시간이 부분적으로 평가에 사용된다 하더라도 기술 향상이라는 직접적인 이점을 얻게 되므로 자신의 시간을 정당하게 사용했다고 인식할 수 있습니다. 형성 평가, 지능형 튜터링 시스템 및 평가와 혼합된 기타 교육 형태는 학습자가 평가에 더 많은 시간을 할애하도록 유도하고, 원칙적으로 기술 측정을 개선하기 위해 이 원칙을 활용합니다.

또 다른 전략은 시험 응시 경험을 더욱 매력적이고 즐겁게 만들어 응시자들이 기꺼이 더 많은 시간을 시험에 할애하도록 하는 것입니다. 멀티미디어 시험, 게임 기반, 게임화된 및 게임 방식으로 설계된 평가(Landers & Sanchez, 2022)는 많은 어린이와 성인이 게임을 즐기고 보상 없이 또는 인지된 외적 직접적 이점 없이 자발적으로 그렇게 한다는 사실을 이용합니다. DARPA의 DARWARS 프로그램은 학생들이 가상 세계에서 시뮬레이션, 지능형 에이전트 및 온라인 커뮤니티를 갖춘 멀티플레이어 게임을 통해 훈련 경험에 자발적으로 수백 시간을 소비함으로써 기술을 습득할 것이라는 아이디어를 기반으로 했습니다 (O'Neil et al., 2004).

대역폭 문제는 일부 상황과 세계 일부 지역에서 존재할 수 있으며, 이는 타당성 위협이 될 수 있습니다.

5.2.2 언제, 어디서나, 보안과 함께

언제 어디서나 시험을 볼 수 있는 방식(Anytime-Anywhere Testing)은 시험 응시자의 편의성과 비용 절감을 위한 요구에 맞춰 오랫동안 연구되어 왔습니다. Bennett(1998)는 이미 전용 시험 센터가 사라질 가능성을 제기했으며, 코로나19 팬데믹이 이러한 변화 속도를 가속화했습니다. 현재는 대규모 시험에서도 언제 어디서나 응시할 수 있는 방식이 현실이 되었으며, 이는 비용 절감, 편의성 증가, 접근성 향상이라는 큰 이점을 가져왔습니다. 그러나 시험 센터는 여전히 존재하며, 많은 사람들이 시험 센터에서 응시하는 것이 더 편리하고 비용이 적게 드는 경우도 있습니다. 특히 기업, 대학, 협회 등에서는 보안 문제, 추가 평가 요소 활용, 응시자 경험 개선, 공정한 디지털 접근성 보장 등의 이유로 대면 시험을 여전히 요구하고 있습니다.

원격 시험(At-home Testing) 또는 모바일 시험(Mobile Testing)은 **고위험 시험(High-Stakes Testing)**에서 보안 문제가 더 복잡해집니다. 현재 보안 문제는 다양한 방식(Choi et al., 2021; ETS, 2023b; Qian et al., 2018a, 2018b)으로 해결되고 있으며, 지속적인 모니터링과 개선이 필요합니다. 한편, 보안이 크게 중요하지 않은 **저위험 형성 평가(Low-Stakes Formative Testing)**에서는 모바일 시험이 많은 장점을 가집니다. 예를 들어, **Karay et al.(2020)**의 연구에 따르면, 모바일 기기에서 시간 제한 없이 시험을 보는 것이 점수에 영향을 주지 않으며, 오히려 다음과 같은 긍정적 효과를 보였습니다. - 학생들이 시험에 더 많은 시간을 할애함. - 책이나 온라인 자료를 적극적으로 활용함. - 결과적으로 학생들에게 더 높은 수용도를 얻음.

5.2.3 새로운 기기

일부 시험은 여전히 종이와 연필 형식으로 시행되지만, 점점 그 빈도가 줄어들고 있다. 심지어 SAT도 2024년부터 완전히 디지털 방식으로 전환되었다(College Board, 2023). 대학원 및 전문대학원 입학을 위한 고위험(high-stakes) 시험들은 2000년대에 디지털 기반 평가로 전환되었으며, 세계 여러 나라를 대상으로 하는 대규모 국내외 평가도 2015년부터 2020년 사이에 디지털 방식으로 전환되었다. 다만 일부 예외가 존재하는데, 예를 들어 OECD의 개발도상국 대상 프로그램인 PISA-D는 여전히 전통적인 시험지 방식으로 시행되고 있다. 초기 디지털 시험 전환은 종이 시험 형식에 비해 큰 기능적 차이를 제공하지 않았으며, 단순히 더 많은 시험 형태를 제공하고 적응형 평가를 가능하게 하는 수준에 그쳤다. 그러나 점차 새로운 기능들이 추가되면서 동영상, 시뮬레이션, 상호작용이 가능한 시험 환경이 도입되었으며, 이러한 추세는 앞으로도 계속될 것으로 보인다. 궁극적으로 교육 평가 분야에서도 더욱 몰입감 있고 흥미로운 형식의 시험이 가능해질 전망이다. 다만, 기술 투자의 수익성이 교육보다 엔터테인먼트 분야에서 훨씬 높기 때문에 이러한 변화는 상대적으로 더디게 진행될 가능성이 크다.

애플 비전 프로(Apple Vision Pro)의 혼합 현실(MR) 헤드셋이나 마이크로소프트의 키넥트(Azure Kinect)와 같은 새로운 기술은 시험의 입력 방식(예: 지시문, 문항 자극, 문항 프롬프트)과 응답 방식(예: 제스처, 물건 잡기, 전신 움직임)에 혁신적인 변화를

가져올 수 있다. 이를 통해 새로운 구인을 새로운 방식으로 측정할 가능성이 열리게 된다. 하지만 이러한 기술들은 시장 변동성에 따라 빠르게 변화하기 때문에, 새로운 유형의 시험을 개발하기 위해 특정 기술에 투자하는 것은 상당한 위험을 수반한다. 예를 들어, 마이크로소프트는 2017년에 키넥트 생산을 중단했으며(Lee, 2023), 한때 주목받았던 사회적 측정 배지(sociometric badge) 기술도(Lederman et al., 2016) 이미 단종된 상태다.

5.3 문항 개발

5.3.1 생성형 AI와 문항 모델을 사용한 자동 문항 생성

문항 개발은 전통적으로 전문가들에 의존해왔으며(Lane 외, 2016이 검토 제공), 따라서 비용이 많이 들고 시간이 오래 걸리는 과정이었습니다. 자동 문항 생성(AIG)은 이 과정을 더 효율적이고 표준화된 방식으로 만들 수 있는 매력적인 대안입니다. 초기 AIG 시도들(Irvine & Kyllonen, 2002)은 표본 문항으로 측정되는 목표 지식, 기술, 능력의 포괄적인 문항 모델을 구축하는 데 초점을 맞추었고, 모델의 핵심 구성요소를 조작하여 많은 유사한 변형을 생성했습니다. 이러한 초기 시도들은 원본과 유사한 고품질 문항을 생성하는 데 효과적이었지만 두 가지 중요한 한계가 있었습니다: 각 문항마다 고유한 모델이 필요했기 때문에 확장이 어려웠고, 목표 구성개념에 맥락을 제공하는 텍스트의 다양성이 제한적이어서 문항들이 비슷하게 보였고, 따라서 독립적인 문항만큼의 정보를 제공하지 못했습니다(Bejar 외, 2002). 생성형 AI는 이러한 한계를 극복하는 데 특히 적합합니다; 여러 문항 유형에 걸쳐 광범위한 텍스트를 생성할 수 있습니다. 따라서 최근의 AIG 접근법들은 많은 다양한 문항 유형에 대한 맥락, 문두, 선택지를 생성하기 위해 LLM을 자주 활용합니다(예: Attali 외, 2022; Chan 외, 2022; Gao 외, 2022; Stowe 외, 2022; Zu 외, 2023). 신중한 문항 모델링과 유능한 LLM의 성공적인 결합은 문항 개발에 자동화를 구현하는 매우 유망한 접근법으로 보입니다. 모든 형태의 글쓰기와 마찬가지로 문항을 작성하는 것은 여러 단계로 구성된 과정입니다. 지금까지 문헌에 소개된 AIG 접근법들은 초기 생성에만 전적으로 초점을 맞추었습니다. 이러한 관점에서, 현재의 AIG 접근법들(LLM 기반 접근법 포함)에 대한 더 정확한 표현은 '자동 문항 초안 작성'일 것입니다. 문항 개발에서 자동화의 잠재력을 완전히 실현하기 위해서는, 전체 과정이 자동으로 생성된 문항 초안을 활용하도록 설계되어야 합니다. 초안 문항들은 정확성, 적절성, 공정성에 대해 검토되어야 하고; 난이도와 변별도를 추정하기 위해 보정되어야 하며; 시험자들에게 도달하기 전에 전달 단위(예: 시험지)로 조립되어야 합니다. ETS와 많은 다른 시험 회사들의 문항 개발 과정은 수십 년 전에 정기적인 간격으로 도착하는 일정한 수의 수동으로 작성된 문항 초안을 수용하도록 설계되었습니다; 이러한 기존 과정들은 문항 개발의 효율성과 규모를 달성하기 위해 자동으로 생성된 많은 수의 초안을 활용하는 데 상당한 병목 현상을 일으킬 수 있습니다. 따라서 초기 생성 능력과 함께 전반적인 문항 개발 과정을 혁신하는 것이 매우 중요합니다.

5.3.2 LLM을 활용한 난이도 모델링

문항의 난이도는 시험지를 구성하고 점수를 결정하는 데 중요한 요소이다. 일반적으로 문항 난이도는 많은 수의 응시자(보통 문항당 500~1,000명)의 실제 응답을 바탕으로 추정된다. 그러나 이 방식은 새로운 문항의 수가 응시자 수보다 훨씬 적다는 전제하에 가능하다. 하지만 효과적인 자동 문항 생성(AIG) 시스템이 즉시 많은 문항을 만들어 낼 수 있다면 이 전제는 성립하지 않으며, 기존 방식으로 난이도를 추정하는 것이 새로운 문항 활용의 큰 장애물이 된다. 따라서 문항 난이도를 예측하는 방법이 대안이 될 수 있다. 과거에는 문항 유형별로 개별적인 예측 모델이 필요했고, 예측 알고리즘의 성능도 제한적이었다. 하지만 대형 언어 모델(LLM)은 문항을 입력하면 난이도를 예측할 수 있는 유연한 모델링 도구로 활용될 수 있다. Zu & Choi (2023a, 2023b)의 연구에서는 오픈소스 LLM을 미세 조정(fine-tuning)하여 문항 난이도를 예측한 결과, 기존 최고 성능의 예측 모델(Loukina et al., 2016)을 뛰어넘었으며, 전문가의 난이도 판단보다 훨씬 정확한 결과를 보였다. 또한 ETS 연구진은 예측된 난이도의 불확실성을 보정하는 방법을 개발하여, 이를 평가 심리학적(psychometric) 분석 과정에 반영하고 있다(Lewis, 2001; Mislevy et al., 1993).

5.4 맥락화와 개인화

OECD의 PISA, PIAAC 또는 SSES와 같은 대규모 국제 평가는 전 세계 여러 국가 및 언어로 시행되며, 국가별 성과를 순위표로 비교하므로 비교 가능성이 핵심입니다. 시험은 영어(또는 프랑스어) 원본을 바탕으로 2단계 과정, 즉 적응 및 번역을 거쳐 준비됩니다 (cApStAn & Halleux, 2019; Hambleton, 2002). 적응 단계에서 해당 국가의 이중 언어 사용자는 특정 개념이 자국 문화에서 의미가 있는지, 그리고 원본 국가에서와 동일하게 해석될 것인지 여부를 나타냅니다. 적응이 무엇을 의미하는지 파악하기 위해 ITC 지침(International Test Commission, 2017) 몇 가지를 고려해 보십시오.

- 관심 집단의 시험 의도된 사용과 관련 없는 문화적, 언어적 차이의 영향을 최소화합니다.
- 적응 과정에서 관련 전문 지식을 갖춘 전문가 선택을 통해 대상 집단의 언어적, 심리적, 문화적 차이를 고려해야 합니다.
- 점수에서 도출된 추론의 타당성에 영향을 미칠 수 있는 시행 절차 및 응답 방식으로 인해 발생하는 문화 및 언어 관련 문제를 최소화하기 위해 시행 자료 및 지침을 준비합니다.

일상적으로 모든 국제 시험에서 요구되는 이러한 종류의 적응은 관련 언어와 문화, 시험 내용 및 시험 원칙에 대한 전문 지식을 필요로 합니다. 이는 언어 및 문화 집단 간의 평가 결과 비교 가능성을 보장하기 위한 비용이 많이 들지만 필수적인 과정입니다. 이러한 작업은 국제적인 대규모 평가 작업에만 국한되지 않습니다. 일반적으로 고용 시험인 시험이 여러 국가에서 지역적으로 사용되는 경우(예: ETS의 이전 Workskills for Job Fit은 18개 언어로 시행되었습니다)와 미국 내의 언어 하위 집단(예: 영어 학습자인 스페인어 사용 시험 응시자를 위한 문화적 적응에 대한 ETS의 캘리포니아 K-12 노력)에서도 이러한 적응이 필요합니다. 미국 내의 문화적 하위 집단에도 유사한 종류의 적응을 적용할 수 있다고 주장할 수 있습니다. 물론, 교차 문화 평가에서 확인된 편향(구조적 편향, 방법 편향 및 문항 편향)은 언어 집단 내의

하위 문화에 대한 시험 점수의 적절한 해석에도 관련이 있습니다(van de Vijver & Poortinga, 2005).

비슷한 종류의 적응은 예를 들어 여러 문화권을 대표하는 여러 국가의 잠재 고객을 위한 제안서나 광고를 준비하는 비즈니스 또는 광고 분야에서도 이루어집니다. AI 운영 분야의 선구자이자 Sageable CEO인 Andi Mann은 팟캐스트에서 이러한 종류의 적응이 곧 AI를 통해 이루어질 것이며, 다양한 문화를 위해 콘텐츠의 맥락을 재구성할 것이라고 제안했습니다(Turchin, 2023). Mann은 사업 제안서나 광고 브로셔를 가져와 대상 문화적 가치에 맞게 조정하는 예를 들었습니다. 예를 들어, 문화적으로 더 적절하고 덜 불쾌하게 만들거나, 격식 있는 어조에서 비격식적인 어조로 바꾸거나, 목표를 재구성하여 문화적으로 더 호환되도록 하는 것입니다. 그는 자동 맥락 재구성이 곧 이미지 크기 조정만큼 쉬워질 것이라고 제안했습니다. Lee et al. (2024)은 개인 맞춤형 마케팅에 LLM을 사용하는 방법을 보여주었습니다.

관련 아이디어는 "특정 개인의 특성/자질 집합을 고려하여 경험이나 상호 작용을 적절하게 조정하는 것"으로 정의되는 개인 맞춤형 학습 문헌(Walkington & Bernacki, 2020)에서 비롯됩니다. 그렇다면 개인 맞춤형 평가는 개인 맞춤형 학습에 대한 평가입니다. (이에 대한 추가 논의는 본 보고서의 피드백 섹션에서 참조하십시오.)

5.4.1 개인 맞춤화 및 맥락화 구현을 위한 LLM 활용

기술과 AI는 이전에는 불가능했던 수준으로 시험 콘텐츠 생성에서 개인 맞춤화 또는 맥락화를 (경제적으로 그리고 대규모로) 가능하게 합니다. ETS Human Progress Study(ETS, 2023a) 응답자의 높은 비율(78%)이 AI가 각 개인 학습자의 필요에 맞게 맞춤화하여 학습 평가를 향상시킬 잠재력이 있다는 데 동의했습니다. 이러한 종류의 개인 맞춤화는 자동 문항 생성의 맥락 내에서 또는 외부에서 수행될 수 있습니다.

미리 만들어진 문항 세트에서 시험 양식을 구성할 때 모든 종류의 적응은 주요 과제가 됩니다. 평가에 개인 맞춤화를 통합하려는 노력은 개인을 위한 콘텐츠의 온라인 적응 또는 실시간 경험을 유사하게 만들 수 있을 만큼 충분히 큰 다양한 문항 풀을 통해 이러한 과제를 극복해야 합니다. 미리 만들어진 콘텐츠에 의존하는 현재의 시험 개발 과정에서는 두 가지 옵션 모두 실현 가능하지 않습니다. 따라서 평가 콘텐츠의 자동 적응은 개인 맞춤형 평가로의 전환을 가속화할 수 있는 매우 영향력 있는 혁신이 될 수 있습니다. LLM은 광범위한 텍스트에 대해 이미 사전 학습되었기 때문에 자동 적응 작업에 매우 적합합니다. 또한 텍스트를 의미 정보가 포함된 숫자 벡터로 인코딩할 수 있으므로 컴퓨터 비전에서 신경 스타일 전송 접근 방식의 성공(Gatys et al., 2016)이 텍스트 영역의 적응 문제에 적용될 수 있음을 시사합니다(Hu et al., 2017; Prabhumoye et al., 2018; Shen et al., 2017; Yang et al., 2018).

그러나 이러한 접근 방식을 적절하게 구현하려면 잠재적인 결과를 신중하게 고려해야 합니다. ETS Human Progress Study(ETS, 2023a)에서 응답자의 71%는 AI가 시스템 내의 의도치 않은 편견과 프로그래밍 결함으로 인해 학습 평가에 부정적인 영향을 미칠 가능성을 우려했습니다. LLM은 사전 학습 샘플에 포함된 편견을 물려받으며 이러한 편견을 결과물에 재현할 수 있습니다. 따라서 순진한 LLM 기반 적응 접근 방식은 기존 편견의 추가적인 확산 및 강화로 이어질 수 있습니다. 그러므로 자동화된 적응의 이점을 기존 편견의 재생산 없이 실현할 수

있도록 적응 결과를 감시하고 통제하는 강력한 메커니즘을 구축하는 것이 중요합니다. 여러 문화와 언어에 대한 대규모 교육 설문 조사를 조정해 온 ETS의 역사와 전문 지식, 그리고 콘텐츠를 생성하는 AI 모델에서 이러한 문제를 완화하기 위해 수행된 작업은 이러한 과제를 성공적으로 해결하는 데 중요한 이점을 제공합니다.

5.5 검사 구성

검사 구성

시험 구성은 시험지에 포함할 문항을 선택하는 과정입니다. 문항 선택은 일반적으로 시험 설계도에 명시된 제약 조건에 따라 이루어집니다 (Davey, 2023; Lane et al., 2016). 이러한 제약 조건은 여러 시험지 간의 비교 가능성을 확보하고, 평가하려는 구인(construct)의 정의에 부합하며, 시험이 측정하고자 하는 내용 영역의 모든 측면을 포괄하고 (구인 과소 대표 방지), 구인과 관련 없는 특징을 최소화하는 데 목적을 둡니다. 포함할 수 있는 제약 조건의 종류는 사실상 무한하지만, 일반적으로 시험 길이, 구인, 내용 (예: 1차, 2차 및 특정 내용), 문항 유형 또는 형식, 독립형 또는 세트 내 문항, 인지 수준 또는 지식 깊이와 관련됩니다 (Davey, 2023). 문항 난이도, 문항 변별도 (측정되는 특질에서 높은 점수를 받은 응시자와 낮은 점수를 받은 응시자가 해당 문항을 정확하게 풀 가능성 정도) 및 문항 완료 예상 시간과 같은 심리측정 속성, 그리고 일반적으로 문항 응답에 영향을 미치는 구인과 무관한 영향을 최소화하기 위해 균형을 맞춘 시험지에서 남학생과 여학생의 언급 횟수와 같은 문항 내용과 관련된 모든 종류의 세부 사항도 포함될 수 있습니다. 심지어 형식 (예: 7페이지 이하, 페이지당 최대 50줄)도 구성 과정의 일부로 포함될 수 있습니다 (Diao & van der Linden, 2013).

Stocking과 Swanson(1993)의 초기 운영 시연과 van der Linden(2005)의 자료집 이후, 자동화된 구성의 이점, 즉 시험 구성을 조합 최적화 문제로 취급하는 것의 이점은 분명해졌습니다 (Davey, 2023). 이는 항공사가 좌석을 채우는 데 사용하는 기술과 군대 및 소매업체가 진열대를 채우는 데 사용하는 기술과 동일합니다. 조합 최적화에서 목적 함수는 제약 조건 집합에 따라 최소화됩니다. 시험 설계도는 제약 조건 집합으로 지정되고, 목적 함수는 목표 평균 문항 난이도 또는 목표 시험 특성 곡선(예: 모든 숙련도 수준에 대한 정보를 제공하는 양식 또는 특정 절단점 주변에 가장 많은 정보를 제공하는 양식) 또는 시험 정보 함수 또는 둘 다를 달성하는 것과 같은 시험 설계를 위한 특정 목표를 달성하도록 선택됩니다 (Ali & van Rijn, 2016). 목적 함수는 내용 목표 또는 보안 목표(예: 문항 노출 최소화; Davey, 2023)를 달성하는 데에도 사용될 수 있습니다.

시험 구성에 대한 조합 최적화 접근 방식은 매우 강력하며 시험의 질적 향상에 기여할 수 있는 기반 역할을 할 수 있습니다. 이 접근 방식은 제약 조건과 목적 함수로 사용할 수 있는 문항에 대한 가용 데이터에 의해서만 제한됩니다. Davey (2023)는 기술 초기 시절의 한 사건을 회상했는데, 자동화된 시험 구성이 한 양식을 만드는 데 사용되었고, 검토하던 시험 개발자들이 그들에게는 명백하지만 알고리즘에는 분명하지 않은 결함을 발견했는데, 그것은 물을 주제로 한 문항이 너무 많았다는 것이었습니다. Davey는 물에 대한 공식적인 내용 요구 사항이 없었기 때문에 알고리즘은 그러한 과다 대표를 알 수 없었고, 이는 인간 검토자들에게 두드러져 보였다고 말했습니다.

물"과 같은 많은 문항 특징이 문항 은행에 포함되지 않는 이유는 이전에 이를 포함해야 할 이유가 없었고 (인간 작성자가 쉽게 발견할 수 있는 것) 관련 있을 수 있는 모든 가능한 특징에 대해 문항을 수동으로 코딩할 시간과 인력이 없었기 때문입니다. 이제 자동화된 구성 접근 방식이 더 널리 사용되고 많은 문항 특징을 쉽게 처리할 수 있게 되었으므로 "물" 문제, 즉 문항을 손으로 코딩하지 않고도 많은 수의 특징별로 분류하는 문제는 새로운 접근 방식에 적합한 영역으로 보입니다.

5.6 보안과 품질 관리

중요한 시험 환경(학교 입학, 장학금 제공, 취업 선발 시험)에서 시험 업계는 고용주 또는 교육 기관이 의사 결정 목적으로 정보를 필요로 하는 학생 또는 지원자의 기술에 대한 신뢰성 있고 타당하며 신뢰할 수 있는 정보를 독점적으로 제공합니다. 이러한 결과는 표준화된 시험 가치 제안의 주요 구성 요소입니다. 개인의 기술에 대한 다른 정보 소스(추천서, 자기소개서, 이력서)는 후보자의 기술에 대한 제한적인 유용한 정보를 제공하며 손상 및 편견에 매우 취약합니다. 자기소개서는 시험 점수와 이전 성적이 고려된 후에는 성적이나 교수진 평가를 예측하지 못합니다 (Murphy et al., 2009). 아마도 자기소개서가 후보자 자신으로부터의 입력뿐만 아니라 (예: 친구, 가족 구성원, 전문가) 다른 출처의 입력을 반영하기 때문일 것입니다 (Powers & Fowles, 1997). 추천서는 자기소개서보다 결과 예측력이 더 높지만 (Kuncel et al., 2014) 자체적인 문제가 있습니다. 부정적인 언급에 대한 강한 편견이 있고, 두 평가자 간의 합의 수준이 작은 경향이 있으며, 추천서 작성자는 평가 심각도에 차이가 있고, 평가에 영향을 미칠 수 있는 평가자의 동기 (예: "내 학생에게 직업을 찾아주겠다" vs. "신뢰할 수 있는 정보 소스라는 내 명성을 유지하겠다")가 투명하지 않습니다. 이력서는 표준화되지 않았으며 후보자의 성별, 인종 및 연령과 같이 기술과 관련 없는 많은 특징을 드러내어 후보자 평가에 편견을 줄 수 있으며 (Kessler et al., 2019) 차별적인 기회를 반영합니다. 이력서와 표준화된 대응물인 바이오 데이터 또한 위조에 취약합니다 (Law et al., 2002).

표준화된 시험은 이러한 대안적인 측정 방법과 관련된 편견과 타협에 덜 취약합니다. Leonhardt(2024)가 지적했듯이 "아마도 시험에 찬성하는 가장 강력한 주장은 입학 과정의 다른 부분들이 훨씬 더 큰 인종적, 경제적 편견을 가지고 있다는 것일 것입니다." Chetty et al.(2023)은 시험 점수가 고등학교 성적보다 결과(엘리트 대학원 진학, 명문 회사 취업)의 더 강력한 예측 변수임을 보여주었습니다. 그들은 또한 상위 1% 소득 학생들의 명문 학교 입학 특혜의 대부분이 시험 점수가 아닌 더 높은 비학업적 평가(세습적 선호도 및 운동 선수 모집과 함께)에서 비롯됨을 보여주었습니다.

그러나 이는 시험 점수의 신뢰성이 보안 절차와 품질 관리를 받는 한에서만 유효합니다. 표준화된 시험을 통해 얻은 점수 정보를 보안상의 허점이나 품질 관리 실패로 인해 신뢰할 수 없다면 표준화된 시험의 가치는 크게 감소합니다.

시험에 대한 잠재적인 보안 위협의 성격은 무엇일까요? 기본적으로 세 가지 위협, 즉 사칭자, 정보 제공자, 그리고 유출된 정답 및 문제 자료가 있지만, 이들은 다양한 형태로 나타납니다. 사칭자는 전통적으로 시험장에서 후보자를 대신하여 시험을 보는 사람이었으며, 이는 원격 시험으로 인해 더 쉬워질 가능성이 있습니다. 정보 제공자는 사칭자와 유사하며, 가정 시험

중에 컴퓨터 카메라 시야 밖에서 방에 숨어 있다가 후보자에게 답을 알려줄 수 있습니다. 또 다른 종류의 정보 제공자는 뛰어난 시험 능력을 보여준 ChatGPT입니다 (Panthier & Gatinel, 2023). 미래의 정보 제공자는 정교한 통신 기술을 통해 후보자에게 정보를 전달하거나 후보자를 대신하여 시험을 보면서 보안 취약점을 악용할 수 있습니다. 마지막 보안 위협은 유출된 정답이며, 이는 전통적으로 전문 시험 응시자 또는 다른 시험 응시자의 시험 문항에 대한 집단 기억으로부터 생성되었습니다. 미래의 유출된 정답은 ChatGPT와 같은 AI 도구를 사용하여 생성될 수 있습니다. 기술을 보여주려는 개인적인 동기와 그러한 증거의 진실성을 보장해야 하는 더 큰 시스템의 필요성이 충돌하는 한, 부정행위와 적발은 고양이와 쥐 게임으로 남을 수 있습니다.

5.6.1 부정행위 탐지와 품질 관리 수행을 위한 접근법

Lee 외(2014)는 대리시험(가짜 응시자), 답 복사(비의도적 정보 제공자), 사전 문제 유출, 집단 공모(정보 제공자)를 탐지하기 위해 설계된 다양한 통계적 부정행위 탐지 방법과 품질 관리 도구에 관한 연구를 검토했습니다. Sinharay(2023)는 Lee 외의 연구에 추가 방법을 보완했습니다. 탐지 방법에는 큰 점수 차이 분석법, 관련된 구성 요소를 측정하는 시험 영역 간 불일치 수행 탐지법, 응답 시간 분석을 통한 개인의 불일치 응답 패턴 파악 등이 포함됩니다. 널리 사용되는 부정행위 탐지 방법은 시험장에서 가까이 앉은 응시자들처럼 '객관식 시험에서 두 응시자의 오답이 비정상적으로 일치하는 경우'(Holland, 1996, p. 2)와 같이 응시자 그룹의 특이한 응답 패턴을 찾는 것입니다. 이는 k-지수와 오답 일치의 본페로니 조정 확률(PMIR; Lewis & Thayer, 1998)을 통해 수행됩니다. 현대의 통신 기술로 인해 응시자들이 서로 가까이 있지 않더라도, 많은 그룹이 유포된 정답을 공유하여 일치하는 응답을 할 수 있습니다. 통계적 방법은 정확히 또는 거의 일치하는 패턴이 비정상적인지를 식별할 수 있으며(Haberman & Lee, 2017), 이를 위한 특허 시스템도 있습니다(Haberman 외, 2022).

시험 과정의 신뢰성을 평가하는 또 다른 방법은 장기적 품질 모니터링입니다(Lee 외, 2014). 누적합 차트는 품질 관리에서 흔히 사용되며 시험에도 적용할 수 있습니다(Lee & Lewis, 2021). 예를 들어, 반복 노출 후 더 이상 이전과 같은 응답을 이끌어내지 못하는 문항을 식별하는 데 도움이 될 수 있는데, 이는 과다 노출을 나타낼 수 있습니다. 더 넓게 보면, 시간에 따른 개별 문항의 급격한 변화를 탐지할 수 있는 다양한 새로운 통계적 방법들이 있습니다. 여기에는 조화 회귀분석(Lee & Haberman, 2013, 2021), 시계열 방법(Lee & von Davier, 2013), 그리고 시험을 포함한 다중 데이터 스트림과 관련된 많은 응용에 적용할 수 있는 순차적 변화 탐지(Chen 외, 2022)가 포함됩니다. 이러한 방법들은 적어도 자주 시행되는 시험에서는 채점과 문항 풀에서 제외해야 할 문제가 있는 문항을 식별하는 데 적용될 수 있습니다. 이러한 새로운 방법들의 추가 발전은 앞서 제안된 새로운 시험 형태들에 대한 적용 가능성을 확장하고, 문항당 응시자 수가 적은 더 큰 문항 풀에 대한 적응을 포함할 수 있습니다.

5.6.2 AI를 활용한 LLM 부정행위 탐지의 새로운 접근법

ChatGPT와 다른 LLM들의 각종 고부담 시험에서의 사용은 부정행위 탐지에 새로운 도전과제를 제시합니다. 음성 복제와 딥페이크를 이용한 부정행위는 새로운 우려를 낳고 있습니다. Hao 외(2024)는 다양한 접근법을 제시했습니다. 여기에는 추가 카메라 사용과, 비판적 사고력과 수행 기반 과제처럼 LLM 도움에 덜 취약한 문항을 포함하도록 시험을 재설계하는 등의 예방 조치가 포함됩니다. 또한 특히 에세이 답안에서 ChatGPT의 기여를 탐지하도록 설계된 탐지기 조치도 포함됩니다. 이러한 탐지기는 ChatGPT를 매우 정확하게 탐지할 수 있지만, 오탐지가 우려됩니다. Hao 외(2024)는 탐지기가 성공적이려면 모든 지표(거짓 양성과 참 음성 비율, 동일 오류율과 대조 표본)를 고려해야 한다고 지적했습니다. 탐지기는 AI 생성 텍스트에 대한 인간의 수정에 강건해야 하고, 하위 집단 편향을 고려해야 하며, 짧은 응답은 구별하기가 더 어렵고, 결국 탐지기는 확률적 증거만 제공할 수 있습니다. 상황은 빠르게 변화하고 있으며, 오픈소스 LLM들은 새로운 도전과제를 제시할 것입니다(Chakraborty 외, 2023; Liu, Zhang 외, 2023; Tang 외, 2023).

5.7 채점 - AI 채점 방법

전통적인 객관식 시험과 그 변형에 대한 채점과 채점 응용은 과학적으로나 운영적으로나 잘 정립되어 있습니다. van der Linden(2018)의 편집본은 보건, 마케팅, 임상심리학, 국제평가 등 다양한 분야의 시험에서 모델링, 분석, 채점, 문항 보정, 개인 및 모델 적합에 사용될 수 있는 다양한 문항반응이론 접근법을 포괄적으로 다룹니다. 모든 종류의 시험에 적용 가능한 접근법을 다루는 다른 많은 연구들도 있습니다(Ostini & Nering, 2006; Wainer & Thissen 2001). 이러한 방법들은 아직 운영이나 연구에서 보편적으로 적용되지는 않고 있으며 - 많은 경우 점수는 단순히 응시자의 정답 수의 합계입니다 - 그러나 조직행동(Lang & Tay, 2021)과 정책 및 보건(Nguyen 외, 2014) 같은 다양한 분야에서 전통적인 합산(고전) 방식을 모델 기반(문항반응이론) 방식이 점차 대체하고 있습니다. 하지만 평가의 미래에서 중요한 부분으로 부상할 것 같은 몇 가지 채점 주제들이 있습니다. 이러한 주제들에는 자동 생성된 문항의 채점, AI 방법을 사용한 에세이 채점, 그리고 새로운 혁신적 문항 유형과 시험 없는 평가의 채점이 포함됩니다.

5.7.1 자동문항생성(AIG)과 문항 난이도 모델링을 위한 채점 방법

자동문항생성에는 여러 접근법이 있습니다(Gierl & Haladyna, 2013의 여러 장 참조; 특히 Irvine & Kyllonen, 2002; Sinharay & Johnson, 2013). 근본요인과 부수요인 접근법은 요인이나 차원들의 집합으로부터 그 차원들의 값을 변화시켜 문항을 구성하는 것을 포함합니다. 난이도에 영향을 미치는 요인들을 근본요인이라 하고, 그렇지 않은 것들을 부수요인이라고 합니다. 이러한 요인들은 영역에 대한 인지적 분석을 기반으로 합니다. 이는 Embretson(1994)과 Kyllonen 외(2019)가 취한 접근법으로, 점진적 행렬이나 수열과 같은 유동적 추론 문항과 같은 알고리즘적 문항에 이상적으로 적합합니다. 이는 데이터를 모델링하고 채점의 기반으로 선형 로지스틱 검사 모형과 그 확장을 사용합니다."

다른 접근법은 문항-모델 접근법으로, '슬롯-채우기'라고도 불리며, 모델 문항의 일부(예: 산술 문장제의 수치들)를 잠재적 채우기 값들의 집합을 가진 슬롯으로 취급합니다. 이는 Bejar 외(2002)와 Graf와 Fife(2012)가 취한 접근법으로, 수학이나 물리 문장제에 이상적으로 적합합니다. Johnson과 Sinharay(2005)는 이러한 접근법들의 채점 방법을 검토하고, '동일 형제' 모델이라 불리는 간단한 모델이 채우기 값에 관계없이 같은 문항 모델에서 만들어진 모든 문항이 동일하다고 가정함으로써 응시자의 능력을 상당히 잘 추정한다고 제안했습니다. 하지만 이 가정을 완화한 '관련 형제' 모델(Glas & van der Linden, 2001)과 '선형 문항 복제' 모델(Geerlings 외, 2011)은 부가 정보의 포함을 허용하고 더 엄격한 통계 분석을 가능하게 함으로써 잠재적으로 훨씬 더 넓은 범위의 자동문항생성 평가에 적용될 수 있습니다.

5.7.2 에세이와 기타 채점하기 어려운 과제의 채점

자동화된 기계 에세이 채점은 이제 운영적 채점에서 잘 정립되어 있습니다. 현재 버전들은 주로 다중 회귀분석과 랜덤 포레스트, 그래디언트 부스팅 머신과 같은 다른 예측적 AI 접근법을 기반으로 한 통계적 학습 방법에 기초합니다(Madnani & Cahill, 2018; Rupp, 2018; Shermis & Burstein, 2013). 자동 에세이 채점은 인간 채점만큼 정확하며, 채점자의 피로도, 엄격성과 관대함, 기준 변화, 시간대, 후광 효과와 관련된 인간의 편향을 피할 수 있는 장점이 있습니다(Williamson 외, 2012). 반면, 자동 채점이 블랙박스이며 자체적인 편향을 가질 수 있다는 인식이 있어 응시자들의 신뢰 부족을 야기할 수 있습니다(Kumar & Boulanger, 2020). 딥러닝 모델과 LLM들이 에세이 채점과 다른 채점하기 어려운 과제의 평가에 사용되기 시작했습니다. 이들은 정확도를 높이고 응시자들의 평가 결과물의 장단점에 대해 더 나은 설명을 제공할 잠재력이 있습니다(Kumar & Boulanger, 2020). Hao 외(2024)는 LLM의 자동 채점 적용 사례들을 논의했습니다. 한 연구는 TIMSS 2019의 여섯 문항에 대해 8개국, 6개 언어의 학생 응답에 대한 인간 평가와 AI 기반 자동 채점 간에 매우 높은 상관관계를 발견했습니다(Jung 외, 2022). 특히 ChatGPT로 번역된 응답으로 시스템을 훈련시켰을 때 관계가 더욱 강했습니다. 다른 연구는 합성곱 신경망을 TIMSS 2019 그래픽 응답 문항 채점에 적용하여 높은 정확도와 인간 평가 편향의 식별을 발견했습니다(von Davier 외, 2023). 이 연구는 유망하지만 초기 단계에 있으며, 앞으로 수년간 단답형, 에세이, 그래픽 응답 및 기타 채점하기 어려운 과제의 채점에 LLM을 적용하는 활발한 활동이 있을 것으로 예상됩니다. 이 작업의 주요 과제는 AI 기반 채점 모델의 편향 회피가 될 것이며, 이는 Duolingo 영어 시험의 책임있는 AI 표준에서 다뤄진 주제입니다(Johnson, 2024). Johnson 외(2022)는 글쓰기 스타일, 응답 길이, 오타와 같이 수행과 관련된 '채점기준 외' 응답 특성이 인구통계학적 변수와도 연관될 수 있다는 예를 논의했습니다. 이러한 문제들에 대한 해결책이 제안되기 시작했으며(Johnson & McCaffrey, 2023), 채점에서의 AI 편향은 유망한 연구 분야로 남을 것 같습니다. 현재 시점에서 LLM 지원 문항 개발과 응답 채점은 LLM의 환각과 AI 편향으로 인해 완전 자율 시스템이 불가능하므로 적극적인 인간 참여가 필요한 연구 주제로 남아있습니다.

5.7.3 무시험 평가의 채점

무시험 평가는 명시적인 시험과 연결되지 않은 행동이나 행동 흔적을 기반으로 한 기술 평가로 정의될 수 있습니다. 여기에는 문제 해결이나 학습 중의 대화, 취업 면접(Emerson 외, 2022), 게임이나 마이크로월드 환경을 자유롭게 탐색할 때 취하는 행동, 심지어 이력서 항목들도 포함됩니다 - Cattell(1965)의 용어로 L 데이터입니다. 이는 서로 다른 활동들의 집합이며, 따라서 이러한 환경에서의 행동을 모델링하기 위해 다양한 접근법이 시도되었습니다. 대부분 이러한 접근법들은 심리측정 문헌과 연결되지 않았습니다. 방법들은 성별과 시간에 따른 학생들의 과제 수행/비수행 행동 패턴 연구(Godwin 외, 2016)부터 탐색적 문항이 있는 표준화 시험에서의 키스트로크 패턴 탐색(He 외, 2019), 비참여의 지표로서 설문 문항 건너뛰기 검토(Hitt 외, 2016; Kyllonen & Kell, 2018; Mignogna 외, 2023), 기계학습 방법을 사용한 대화 코딩 특성화(Kyllonen 외, 2023)까지 다양합니다. 이 분야는 분류와 다른 종류의 데이터 탐색을 위해 LLM 접근법을 사용하는 중요한 추가 발전이 있을 것 같습니다.

5.8 공정성

공정성, 즉 편향의 최소화는 시험 점수 해석의 정당성이나 타당성에 영향을 미치기 때문에 시험에서 '최우선적이고 근본적인 관심사'로 여겨집니다(AERA 외, 2014, p. 49). 시험 점수 해석자는 장애 여부, 언어 상태, 문화적 또는 언어적 배경과 같은 응시자의 특성에 관계없이 시험이 동일한 기저 구인을 측정한다고 가정할 수 있어야 합니다.

표준의 의미 내에서 공정한 시험은 모든 응시자에 대해 동일한 구인(들)을 반영하며, 의도된 모집단의 모든 개인에 대해 동일한 의미를 가진 점수를 제공합니다. 공정한 시험은 의도된 구인과 무관한 특성으로 인해 특정 개인들에게 이점이나 불이익을 주지 않습니다. ...의도된 모집단의 모든 개인의 특성(인종, 민족, 성별, 연령, 사회경제적 지위, 언어적 또는 문화적 배경과 관련된 특성 포함)은 공정한 평가에 대한 장벽을 줄일 수 있도록 개발, 시행, 채점, 해석 및 활용의 모든 단계에서 고려되어야 합니다.(AERA, 2014, p. 50)

시험 공정성에 대한 이러한 개념은 문항 작성 단계에서 '제품이나 서비스의 목적을 충족하기 위해 필요한 경우를 제외하고는 일반적으로 성차별적, 인종차별적, 또는 불쾌감을 주는 것으로 여겨지는 상징, 언어, 내용을 제거하도록 설계된' 지침들을 고려함으로써(ETS, 2014, p. 21; ETS, 2022 참조), 그리고 문항 내용의 접근성과 공정성을 점검하는 검토 과정을 통해 다룰 수 있습니다. 공정성은 또한 성별, 인종, 언어, 문화 및 기타 요인에 기반한 서로 다른 집단에 대해 시험이 동일한 구인을 측정하는 정도를 조사하는 문항 반응의 통계적 분석을 통해서도 다뤄집니다. 통계적 방법은 예를 들어 성별 집단 간 단어에 대한 차별적 친숙도(예: 스포츠 용어) 또는 문화 집단 간 차이(예: 음식 항목) 때문에 두 집단에서 같은 방식으로 작동하지 않는 문항을 식별하는 데 사용될 수 있습니다. 이 주제에 대한 논의는 Millsap(2011)에서 찾을 수 있습니다.

공정성의 두 번째 정의는 성별, 인종, 연령으로 정의된 서로 다른 응시자 집단의 선발률에 기반한 채용 시험에서의 우려사항입니다. 만약 선발 절차가 부정적 영향을 미쳐 보호

집단 구성원들을 가장 선호되는 집단보다 더 높은 비율로 걸러낸다면, 고용주는 직원 선발 절차에 관한 통일 지침을 위반할 수 있으며(EU에서는 간접 차별이라는 유사한 개념), 평등고용기회위원회의 법적 집행 조치를 받을 수 있습니다.

Bennett(2023)는 Solano-Flores(2019)와 Sireci(2020)의 견해에 따라, 이러한 정의들을 넘어서 '시험이 우리가 빠르게 되어가고 있는 다원적 사회에 더 이상 적합하지 않은 세계관을 대표한다'는 인식(pp. 17-18) 때문에 전통적인 표준화 시험에 대한 반대와 관련하여 교육평가의 기본 전제를 재고해야 한다고 주장했습니다. 그의 제안은 내용을 문화적으로 관련성 있게 변경하고, 인구 특정적 평가를 제공하며, 학생 특성에 맞게 평가를 조정하고, 학습자 주체성을 장려함으로써 '사회문화적 반응형' 평가(CRA)를 설계하는 것이었습니다(O'Dwyer 외, 2023).Bennett는 시험의 문화적으로 관련된 문제들이 응시자들의 평가에 대한 동일시, 참여와 동기부여, 사전 지식의 활성화, 그리고 결과적으로 그들의 시험 수행과 자신감, 효능감을 증가시킬 것이라고 가정했습니다. 문화간 연구의 관점에서(이 보고서의 맥락화와 개인화 섹션 참조), Bennett는 일반적인 적응보다 더 광범위할 수 있는 적응 단계를 평가에 제안했습니다. Walker 외(2023)는 '신념, 가치관, 윤리; 그들의 삶의 경험; 그리고 그들이 어떻게 배우고 행동하고 소통하는지에 영향을 미치는 모든 것'과 같은 학생들의 배경 특성을 고려하는 CRA 설계를 위한 잠정적 원칙을 제안했습니다(p. 1). Dobrescu 외(2021)와 Kukea Shultz와 Englert(2021)는 CRA를 현장 테스트했지만, 둘 다 CRA가 비CRA 버전의 시험과 동등하다는 것을 공식적으로 입증하지는 않았습니다.

Sinharay와 Johnson(2023)은 이러한 한계를 다루고, '표면적으로 동등하지 않은 대안적 과제 형태로부터 응시자에 대한 동등한 증거를 얻는' CRA의 데이터를 분석하기 위한 통계적, 심리측정적 프레임워크를 제안했습니다(Mislevy, 2018; 동등성 논의는 Feuer 외, 1999 참조). Sinharay와 Johnson(2023)은 기준 집단용(RGV)과 Bennett(2023)에 따라 초점 집단용으로 수정된(FGV) 두 가지 형태의 문항을 짝지어 이를 달성했습니다. 연구자는 전문가 판단과 형태 내 심리측정 분석(난이도, 변별도, 신뢰도, 요인 구조, 차별적 문항 기능[DIF], 검사 특성 곡선)을 통해 형태 간 동등성을 확립합니다. 이를 통해 연구자는 정책적으로 호환 가능한 형태별 점수를 산출합니다. 하지만 응시자들은 두 형태(RGV, FGV) 모두에서 점수를 받을 수도 있으며, 이는 맥락 내와 맥락 외 능력을 측정합니다. 어느 집단이 어떤 형태를 받는지와 관련된 다양한 설계를 테스트하는 시뮬레이션 연구에서, Sinharay와 Johnson(2023)은 일부 문항이 형태 간에 본질적으로 공통적이기만 하다면 두 형태의 점수를 비교 가능한 것으로 취급할 수 있을 것이라는 점을 발견했습니다.

응시자의 문화적 배경과 관련된 시험의 편향성이라는 일반적 문제를 다루기 위해 완전히 다른 형태를 만드는 것 외에도 다른 접근법이 가능할 수 있습니다. 예를 들어, De Boeck와 Cho(2021, p. 712)는 개인과 문항 효과를 고정 효과가 아닌 무선 효과로 취급하고, '문항 응답을 이해하는 데 도움이 된다면 문항의 하위 집합과 심지어 전체 시험에 걸쳐 DIF가 퍼지는 것을 허용하면서 변동을 설명하기 위해' 설명 공변량을 사용하는 통계적 개념에 기반한 대안적 DIF 범주를 제시했습니다. De Boeck(2023)은 자극 자료에 대한 친숙도가 다양한 참가자들의 예를 사용했는데, 이러한 변동이 수행과 관련이 있었습니다(즉, 설명 공변량). 문화적 친숙도나 학습 기회와 같은 이러한 공변량이 조작적으로 정의된다면, 마찬가지로 문항 응답을 설명하는 데 도움이 되는 설명 공변량으로 사용될 수 있고, 문화적 친숙도나 학습 기회를 고려한 시험 채점의 기반으로 사용될 수 있을 것입니다.

5.9 결론: 운영 혁신

시험의 목적, 관리 조건과 제약사항에 대한 고려와 함께 문항 개발, 시험 구성, 보안, 품질 관리, 채점, 시험 평가를 포함하는 시험 운영은 시험 산업의 핵심입니다. 시험을 타당하고, 신뢰할 수 있으며, 공정하고, 응시자와 다른 이해관계자들에게 유용하게 만드는 것과 관련된 운영에는 많은 도전적인 문제들이 있습니다. 시험의 시작부터 그래왔듯이, 특히 LLM과 다른 AI 기술을 포함한 기술의 발전이 시험 운영에 극적인 영향을 미칠 것 같습니다. 우리는 시험이 어떻게 개발, 구성, 채점되는지, 어떻게 보안이 유지되는지, 그리고 모든 응시자가 시험의 가치를 인식하고 시험 점수를 기반으로 한 추론이 적절하고 정당하다고 확신할 수 있도록 공정성이 확보되는 방식과 관련하여 효율성과 품질에서 상당한 발전을 볼 것 같습니다.

6 피드백: 학습과학 기반의 시험 응시자를 위한 통찰과 실행 계획

"과거에는 모든 사람이 동일한 기준에 따라 평가받았습니다. 하지만 미래에는 개인의 능력과 목표에 기반한 맞춤형 평가를 개발할 수 있을 것입니다. 이는 평가의 큰 발전을 의미합니다."
— 조아나 렌코바 (미래학자, 전략가, Futures Forward)

이 절에서는 평가가 학습을 어떻게 촉진할 수 있는지에 대한 다양한 연구를 검토합니다. 여기에는 평가와 학습의 결합, 형성평가, 테스트 효과, 개별 지도 및 지능형 튜터링 시스템과 관련된 연구가 포함됩니다. 또한 진단 평가, 과정 분석, 효과적인 방법에 대한 기존 연구, 피드백의 영향 등을 살펴보고, 학습 원리에 대한 논의로 마무리합니다. 이러한 모든 연구는 평가를 통해 수험생에게 유용한 정보를 제공하고, 그들이 교육 및 진로 목표를 효과적으로 달성할 수 있도록 지원하는 방법에 대한 시사점을 제공합니다.

Figure 6.1: 능력 평가로 인한 다양한 정서의 증가 가능성을 보고한 응답자 비율

주: ETS 인간 진보 연구(ETS, 2023a)의 데이터. 설문 문항: '능력 평가를 받고 성장을 위한 지침을 받을 수 있다면, 다음 각각을 느끼거나 실행할 가능성이 더 높아지거나 낮아질 것입니까? (덜 가능함/변화 없음/더 가능함)'

6.1 논의의 기초 마련

시험 응시자는 단순히 합격 여부, 입학 여부, 혹은 수상 여부를 넘어 더 많은 정보를 얻기를 원하고 있습니다. ETS의 Human Progress Study (ETS, 2023a)에 따르면, 진로 상담이 포함된 평가를 제공받을 경우 응시자들은 새로운 기술을 배우려는 동기가 더 커지고, 도전에 대비할 수 있으며, 자신의 성과를 인정받고 있다고 느끼고, 자신의 역량과 새로운 직업 기회를 추구하는 데 대한 자신감을 갖는 등 긍정적인 반응을 보였습니다(그림 5 참조).

많은 경우, 시험 점수는 응시자가 자신의 성적을 해석할 수 있도록 기준 점수, 비교 집단 정보, 그리고 기술적 설명과 함께 제공됩니다. 그러나 응시자가 교육 및 진로 목표를 달성하는 데 실질적으로 도움이 될 수 있는 피드백을 제공하기 위해 더 많은 노력이 필요합니다. 이 절의 목적은 학습 과학 원리와 형성 평가, 학습을 위한 평가, 테스트 효과, 피드백, 개별 지도와 관련된 연구 결과를 바탕으로, 응시자에게 실질적이고 실행 가능한 정보를 제공하는 방안을 탐색하는 데 있습니다.

진단 평가는 오랫동안 응시자의 지식과 기술에 대한 깊이 있는 통찰을 제공하고, 보다 맞춤형 피드백의 기반이 될 것으로 기대되어 왔습니다. 그러나 진단 평가의 심리측정 모형이 상당한 발전을 이루었음에도(Rupp et al., 2010), 실제로 그 기대를 완전히 충족시키지는 못했습니다.

진단 평가의 개선과 응시자에게 보다 유용한 정보를 제공하는 데 기여할 수 있는 방법 중 하나는 과정 분석(process analysis)입니다. 과정 분석은 단순히 문항 응답 결과를 해석하는 것을 넘어, 응답 시간, 특정 응시자 행동, 그리고 협력적 문제 해결이나 협력 학습의 경우 문제 해결 과정에서 이루어지는 대화와 같은 추가 정보를 분석하는 기법입니다. 이러한 과정 분석은 학습자가 무엇을 알고 수행할 수 있는지, 혹은 무엇을 모르는지에 대한 유용한 통찰을 제공할 수 있는 원천 데이터를 제공합니다. 또한, 과정 데이터를 진단 모형과 결합하면 응시자의 지식과 개념 이해 수준에 대한 보다 구체적인 해석을 가능하게 할 수도 있습니다.

지능형 교수 시스템(Intelligent Tutoring) 또는 적응형 훈련(Adaptive Training)은 과정 분석이 활용되는 대표적인 분야 중 하나입니다(Greif et al., 2017). 학습자의 행동이 기록된 로그 파일(과정 데이터)을 실시간으로 분석하여 학습자의 지식 상태를 동적으로 모델링하고, 이를 바탕으로 적절한 교수 전략을 선택하며 학습자의 숙달도를 추정합니다. 과정 데이터와 응답 데이터는 학습자의 지식 평가에 기여하며, 학습 전반에 걸쳐 피드백이 제공됩니다.

인간 교사가 학생의 이해도를 질문을 통해 확인하고 이에 맞춰 수업을 조정하는 것과 유사하게, 지능형 교수 시스템도 학습자의 반응을 분석하여 적절한 피드백을 제공합니다. 인간 교사(Nickow et al., 2020; VanLehn, 2011)와 컴퓨터 기반 튜터링 시스템(Duolingo Team, 2023; Sottilare et al., 2018)에서 피드백이 활용되는 방식을 분석하면, 평가 환경에서 피드백을 효과적으로 활용하는 방안을 모색하는 데 도움이 될 수 있습니다.

결론적으로, 평가에서 유용한 피드백을 제공하기 위한 노력은 학습 원리에 기반해야 합니다. 피드백은 교수의 한 형태이며, 효과적인 학습 방법에 대한 기존 연구를 검토하는 것이 중요합니다. 이러한 연구는 학습 원리의 형태로 정리되어 있으며, 이를 평가와 피드백 설계에 반영하는 것이 필요합니다.

6.2 평가와 학습의 결합을 위한 패러다임

이 섹션은 다음과 같은 순서로 구성됩니다. 먼저, 평가와 학습을 결합하는 다양한 패러다임을 살펴봅니다. 여기에는 형성 평가(formative assessment), 이와 거의 동일한 개념인 학습을 위한 평가(assessment for learning), 인지심리학의 기억 연구에서 비롯된 테스트 효과(testing effect), 그리고 인간 및 기계 기반 교수(tutoring)가 포함됩니다. 다음으로, 교수 과정에서 피드백이 미치는 영향을 분석한 연구들을 검토합니다. 이후, 진단 평가(diagnostic assessment)와 과정 분석(process analysis)에 대한 논의를 진행한 후, 학습 원리에 대한 검토를 수행합니다. 마지막으로, 시험 과정에서 제공되는 피드백이 개인의 학습 성과를 향상시키고, 교육적 형평성을 증진하며, 궁극적으로 사회 전체에 긍정적인 영향을 미칠 수 있는 방안을 논의하며 결론을 맺습니다.

6.2.1 형성평가(학습을 위한 평가)

평가가 학습을 향상시키는 방법에 대한 개념과 연구는 매우 다양하다. 그중 하나가 형성 평가(formative assessment)이며, 이는 다양한 정의를 포함하는 폭넓은 개념이다(Bennett, 2011). Xuan et al.(2022, 부록 A)은 여러 연구에서 제시된 19개의 정의를 수집하여 **무엇(what), 왜(why), 언제(when), 누구(who), 어떻게(how)**라는 질문을 기준으로 체계화하였다.

Xuan et al.(2022)의 정의를 비공식적으로 요약하면, 형성 평가는 학습자의 현재 위치와 목표를 파악하고, 학습 능력을 향상시키거나 교수 방식을 조정하는 과정(혹은 도구)이며, 수업 중(혹은 교수 과정에서) 교사(혹은 학생 또는 동료)가 다양한 접근법을 활용하여 수행하는 평가이다.

Shepard(2017)는 형성 평가를 "수업 중 교수와 학습을 개선하는 평가"(p.279)라고 가장 간결하게 정의하였다. 반면, Black & Wiliam(1998)의 정의는 가장 영향력이 크며, "형성 평가는 학생의 필요에 맞춰 교수와 학습을 조정할 수 있도록 정보를 제공하는 평가 활동"(pp. 7-8)이라고 설명한다.

형성 평가와 관련된 개념으로는 학습을 위한 평가(Assessment for Learning, AfL), 학습 자체로서의 평가(Assessment as Learning), 형성적 평가(Formative Evaluation), 교육과정 기반 평가(Curriculum-Based Assessment) 등이 있다. 하지만, Xuan et al.(2022)과 Klute et al.(2017)의 메타분석에서는 이러한 개념들을 본질적으로 구분되지 않는 것으로 보았다. 특히, Xuan et al.(2022)은 형성 평가의 메타분석에서 피드백 개념도 포함하였다. 학습을 위한 평가(AfL)에 대한 비교 가능한 정의 목록이 존재하지 않기 때문에, 두 개념을 동의어로 간주하는 것이 편리하다.

형성 평가의 효과를 분석한 여러 메타분석 연구가 진행되었으며, 다양한 정의로 인해 효과 크기(Effect Size)에 대한 추정치도 상이하게 나타났다.

Fuchs & Fuchs(1986)는 형성 평가의 효과 크기를 0.7로 보고했지만, Kingston & Nash(2011)는 연구 포함 기준의 엄격성 차이로 인해 **영어(0.32), 수학(0.17), 과학(0.09)**에서 더 낮은 효과 크기를 제시했다.

Klute et al.(2017)의 연구에서는 **수학(0.36), 읽기(0.22), 쓰기(0.21)**의 효과 크기를 보고했으며, 형성 평가가 학생 주도(student-directed)인지 교사 주도(other-directed)인지에 따라 차이가 있음을 발견했다.

학생 주도(student-directed) 평가: 학생들이 교사 없이 정해진 절차에 따라 그룹 활동을 수행 (수학 효과 크기: 0.45) 교사 주도(other-directed) 평가: 교사가 직접 지도하고 수업을 조정 (수학 효과 크기: 0.30) 읽기에서는 교사 주도 평가가 학생 주도 평가보다 더 높은 학습 향상을 보임. 그러나 학생 주도 vs. 교사 주도 평가를 비교한 연구 수가 적고, 연구별 개입 방식이 달랐기 때문에 이러한 차이가 학생 주도/교사 주도 여부 때문인지 여부는 명확하지 않다.

Xuan et al.(2022)의 메타분석에서는 추가적으로 다음과 같은 결과를 보고했다.

교사-학생 협력 형성 평가가 교사 주도 형성 평가보다 효과적 **맞춤형 지도(differentiated instruction, 평가 결과에 따라 교수법을 조정)**가 **비맞춤형 지도(nondifferentiated instruction)**보다 효과적 앵글로권(Anglophone)과 유교 전통(Confucian-heritage) 문화권 간 차이가 존재하며, 유교 전통 문화권에서 더 높은 효과 크기가 나타남

6.2.2 검사 효과

검사 효과는 검사를 받는 것의 기억에 대한 이점을 설명하는 용어입니다—특히 개념에 대해 검사를 받는 것이 그 개념의 학습을 향상시킬 수 있다는 것입니다. 검사 효과는 인지심리학의 인간 기억 문헌에서 나왔습니다(Karpicke & Blunt, 2011). 기본 아이디어는 학습이 초기 교수(노출), 이어서 학습(또는 연습), 그리고 최종 검사로 나눠질 수 있다는 것입니다. 실증적 발견은 일부 중간 검사가 일부 학습을 대체하면, 최종 검사가 대체 없이보다 자료에 대한 더 큰 기억을 보여줄 것이라는 것입니다. 이 발견은 비교 조건인 학습 단계가 단순 시연보다 더 큰 기억 향상을 산출하는 것으로 알려진 기억 정교화와 같은 능동적 학습을 포함하는 경우에도 사실입니다. 검사 자체가 학습에 비해 왜 향상을 산출하는지에 대해 생각하는 한 가지 방법은 중간 검사가 특히, 하지만 중간(및 최종) 검사가 회상 검사인 경우에만은 아닌, 인출 연습의 기회를 제공한다는 것입니다. 인출 연습은 나중 검사가 인출을 포함하기 때문에 나중 검사 동안 가치가 있습니다. 따라서 검사 효과 현상의 또 다른 용어는 연습 검사 또는 검사 연습으로, 학습자가 검사를 볼 때 하는 것이 검사 응시를 연습하는 것이라는 아이디어를 전달합니다(Adesope et al., 2017).

Bangert-Drowns et al.(1991)의 초기 연구 이후 검사 효과에 대한 발견을 지지하는 여러 메타분석이 있었습니다. 실험실 연구에 초점을 맞춘 Rowland(2014)는 검사 효과의 다양한 이론적 설명에 대한 증거를 조사했습니다. 그는 재학습과 비교했을 때 검사 효과에 대해 0.50의 효과 크기를 발견했습니다. 또한 검사 효과가 회상에서 더 크지만 재인 검사에서도 여전히 존재하며, 단기와 장기 간격 모두에서 작동하고, 언어적 및 비언어적 자료 모두에서 작동한다는 것을 발견했습니다. 검사 효과는 실험실 연구에 국한되지 않습니다. Phelps(2019)의 메타분석은 검사 효과를 훨씬 더 광범위하게 정의하여 지난 세기 동안 일반적으로 검사의 효과에 대해 수행된 많은 연구를 포함했으며, 0.55에서 0.88에 이르는 효과 크기를 발견했습니다. Adesope et al.(2017)은 Phelps에 비해 범위를 제한하여 양적, 저부담 연구만을 포함했으며, 118개 실험에서 272개의 효과를 조사하여 0.61의 평균 효과 크기를 발견했습니다(학습 자체와 비교했을 때 0.51; 통제 조건이 검사와 무관할 때 0.93).

또한 처치로서 객관식 검사(0.70)가 단답형 검사(0.48)보다 더 큰 향상을 주었고 둘을 함께 했을 때 더 높았으며(0.80), 하나보다 더 많은 연습 검사보다 단일 연습 검사가 가장 좋았고, 효과는 실험실과 교실 환경 그리고 초등, 중등, 고등 환경에서 유사하게 발생했다는 것을 발견했습니다.

검사 효과 외에도, Roediger et al. (2011)은 시험이 교육에 미치는 실질적인 이점들을 확인했습니다. 직접적인 이점은 회상 연습이 학습한 내용을 더 잘 기억하게 하고(테스팅 효과), 관련된 자료까지 기억하는 데 도움을 주며, 새로운 상황에 적용할 수 있도록 도와준다는 것입니다. 개방형 평가 또한 학생들이 정보를 조직하는 데 도움을 줍니다. 간접적인 이점으로는 자주 시험을 보면 학생들이 더 많이 공부하게 되고, 자신의 지식에서 부족한 부분을 발견할 수 있게 됩니다(이것은 명시적 또는 암시적인 피드백, 특히 시험 결과를 알게 되어 발생합니다). 또한 더 어려운 부분에 집중하게 됩니다. Roediger et al.는 자기 시험과 자주 퀴즈를 풀기를 권장했습니다.

6.2.3 튜터링

개인 튜터링, 즉 1:1 또는 소규모 그룹(5명 이하)은 가장 효과적인 교육 방법 중 하나로 간주됩니다. Bloom(1984)은 세 가지 연구에서 얻은 증거를 통해, 좋은 튜터의 1:1 튜터링이 전통적인 교육 방식보다 2 표준편차만큼 향상된 성과(즉, 효과 크기 2.0)를 보였다고 발표했습니다(그리고 형성 평가가 포함된 숙련도 학습보다 약 절반 정도의 향상). 그는 1:1 튜터링이 너무 비용이 많이 든다고 주장하면서, 사회적 목표는 튜터링의 이점을 더 실용적이고 현실적인 방법으로 달성하는 방법을 찾는 것이라고 언급했습니다. 이를 그는 "2시그마 문제"라고 불렀습니다.

Dietrichson et al. (2017)의 메타 분석에서는 36개의 연구를 통해 튜터링이 피드백, 진척도 모니터링, 협력적 학습과 함께 가장 강력한 학업 개입 방법으로 나타났다고 결론지었습니다. 이들 개입은 모두 약간 낮은 효과 크기를 보였지만, 여전히 표준화된 성취도 시험 점수에 미치는 영향에서 가장 큰 영향을 미친 것으로 평가되었습니다(이 연구에서는 14개 개입 유형 중). 특히 저소득층 학생들을 대상으로 한 연구에서 튜터링의 효과 크기는 .36으로 더 겸손하지만 여전히 상당한 효과가 있었습니다(피드백의 경우 .32, 협력적 학습은 .22). 이는 읽기와 수학에서 평균 개입 효과 크기인 .09와 .08에 비해 훨씬 큰 차이를 보입니다. Dietrichson et al.은 연구의 엄격성을 검토했으며(예: 치료 대 통제군 설계, 대부분은 76%가 무작위 통제 시험), 또한 표준화된 성취도 시험을 결과로 사용했는데, 이는 개입 내용의 오염 편향을 피하기 위한 조치였습니다. 이로 인해 **Bloom(1984)**의 튜터링 효과 크기 추정과 **Dietrichson et al.**의 추정치 간 차이가 일부 설명될 수 있습니다.

Dietrichson et al. (2017)은 개입 방법에 초점을 맞춘 반면, Nickow et al. (2020)은 튜터링 자체에 초점을 맞추어 96개의 연구를 검토하고 프로그램 특성과 맥락이 미치는 영향을 조사했습니다. 이들은 효과 크기 추정치를 .37로 제시하며, **"튜터링 프로그램은 PreK-12 수준에서 가장 유연하고 잠재적으로 혁신적인 학습 프로그램 유형"**이라고 결론지었습니다. 연구에서는 튜터링을 교사나 교사 보조원이 진행할 때, 부모가 진행할 때보다 효과가 더 컸으며, 초등학교 저학년에서 더 큰 효과가 나타났고, 학교에서 진행되는 튜터링이 방과 후에

진행되는 튜터링보다 효과적이었다고 밝혔습니다. 또한, 연구팀은 방과 후 부모 튜터링이 실행의 통제가 더 어려운 방식이라고 제안했습니다.

튜터링은 지금까지 확인된 가장 강력한 교육 개입 방법인데, 왜 그렇게 효과적일까요? Nickow et al. (2020)은 여러 가지 가능성을 제시했습니다. 첫 번째 가능성은 튜터링이 일반적으로 교실 수업을 보완하는 방식으로 사용되기 때문에 학습 시간을 더 많이 제공하기 때문입니다. 두 번째는 튜터링이 학생의 수준에 맞는 맞춤형 학습을 제공한다는 점으로, 이는 추적이나 학급 규모 축소를 통해 어느 정도는 구현될 수 있습니다. 세 번째는 튜터링이 학생의 참여를 촉진하고 빠른 피드백을 가능하게 하여, 학생이 더 많은 노력을 기울이도록 자극한다는 것입니다. 마지막으로, 튜터와의 인간적인 연결이나 멘토십 관계가 중요한 역할을 할 수 있습니다.

VanLehn (2011)은 튜터링이 학습 결과에 영향을 미칠 수 있는 잠재적인 메커니즘을 제시하면서, 인간 튜터가 컴퓨터 튜터보다 더 잘할 수 있는 부분에 집중했습니다. 그가 충분한 연구 지원을 찾지 못한 가능한 가설들은 다음과 같습니다:

인간 튜터는 학생의 지식과 오해에 대한 자세한 진단 모델을 개발한다고 여겨지지만, 실제로 인간 튜터가 이를 수행한다는 경험적 증거는 거의 없습니다. 튜터는 학생에게 필요한 정확한 과제를 선택한다고 여겨지지만, 이는 컴퓨터 튜터도 마찬가지로 할 수 있기 때문에 인간의 장점이라고 할 수 없습니다. 인간 튜터는 정교한 튜터링 전략을 사용할 수 있다고 생각되지만, 연구에 따르면 인간은 실제로 복잡한 튜터링 전략을 잘 사용하지 않는다고 합니다. 인간은 주제에 대해 깊은 지식을 가지고 있어 관련 아이디어를 제공할 수 있다고 여겨지지만, 관련 지식이 제공되더라도 결과에 큰 영향을 미치지 않는다고 연구에서 밝혀졌습니다. **Nickow et al. (2020)**와 마찬가지로 VanLehn은 "따뜻한 몸" 효과, 즉 튜터가 칭찬을 통해 동기 부여를 높일 수 있다는 가설을 제시했지만, 이 또한 연구에서 충분한 지지를 받지 못했습니다. 하지만 VanLehn (2011)은 다음과 같은 가설들에 대해 지지를 받았습니다:

인간 튜터는 필요한 순간 즉시 피드백과 힌트를 제공한다. 인간 튜터는 학생의 사고 과정을 돕는 스캐폴딩을 제공한다 (즉, "유도된 프롬프트"를 제공한다). 튜터는 학생이 더 능동적이고 건설적인 행동을 하도록 격려하여 학습을 촉진한다.

VanLehn (2011)은 또한 경험적 문헌에서 지지를 받은 이 세 가지 가설이 Chi와 Wylie (2014)의 상호작용적, 구성적, 능동적, 수동적 (ICAP) 프레임워크와 일치한다고 제안했습니다. 이 프레임워크는 학생들의 참여 행동을 네 가지 ICAP 모드(상호작용적, 구성적, 능동적, 수동적)로 분류할 수 있으며, 학습은 학생들이 학습 자료와 더 많이, 그리고 더 적극적으로 참여할수록 증가한다고 설명합니다. 상호작용적 학습은 참여의 최고 단계로, 그만큼 학습 효과도 가장 큽니다.

6.2.4 지능형 교수(적응형 교수) 시스템

지능형 튜터링 시스템(ITS) 또는 **적응형 교육 시스템(AIS)**은 컴퓨터를 튜터로 활용하는 방법으로, Bloom(1984)의 2-시그마 문제를 해결하려는 노력의 일환입니다. ITS에 관한 방대한 문헌이 존재하며, 11권의 시리즈(및 분석)에 포함된 연구들이 있습니다(Sinatra et al., 2023). 이에는 ITS의 모든 측면에 대한 강점-약점-기회-위협 분석도 포함되어

있습니다(Goldberg & Sinatra, 2023). 전통적인 ITS의 구조는 학습자 모델, 도메인 또는 커리큘럼 모델, 교육학적 모델로 구성됩니다. 학습자 모델은 학습자의 현재 지식과 기술 수준, 그리고 현재 상태를 나타냅니다. 도메인 모델은 가르칠 커리큘럼이나 지침을 나타내며, 도메인 콘텐츠 선택을 위한 규칙(적응적 순서 지정)을 포함합니다. 교육학적 모델은 학습자의 수행에 따라 언제 피드백이 필요한지를 식별합니다(적응적 피드백), 이는 ITS 아키텍처에서 중요한 요소입니다. 더 간단하게 정의하자면, ITS는 문제 해결 중에 개인화된 프롬프트, 힌트, 지원 피드백을 제공하는 시스템입니다(VanLehn, 2011).

적응성이 중요한 요소라는 일부 증거가 있는 것으로 보입니다. 적응형 교육 시스템은 학습자 모델을 사용하여 개인화를 구현합니다(예: 적응형 피드백, 과제나 활동의 적응형 순서). 이 학습자 모델에는 학습자의 인지적, 메타인지적, 정서적, 성격적, 사회적, 지각적 특성에 대한 정보가 포함될 수 있습니다(Abyaa et al., 2019; Shute & Zapata-Rivera, 2012). 학습자 모델은 또한 학습자, 교사 및 다른 대상에게 제공되어 메타인지적 과정, 협력, 내비게이션, 신뢰 및 모델의 정확성을 지원할 수 있습니다(Bull & Kay, 2016). 학습자 모델 정보를 공유하는 데 사용되는 정보 유형과 메커니즘은 각 대상의 필요, 지식 및 태도에 따라 달라집니다(Zapata-Rivera & Forsyth, 2022; Zapata-Rivera, Graesser et al., 2020).

"입력 기회 간 참가자에게 요구되는 추론의 양"을 **곡물 크기(grain size)**로 정의할 수 있습니다(VanLehn, 2011, p. 202). VanLehn(2011)은 피드백이 답변 후에만 제공되는 답변 기반 튜터링(적응형 테스트와 유사)에서부터, 문제 해결 단계 후에 피드백이 제공되는 단계 기반 튜터링, 문제 해결 단계보다 더 세분화된 피드백과 스캐폴딩을 제공하는 서브단계 기반 튜터링에 이르기까지 상호작용의 크기(세분화 수준)가 점진적으로 커지는 연속체를 제안했습니다. 인간 튜터링은 언제든지 개입할 수 있습니다. VanLehn은 상호작용의 세분화 가설을 제안하며, 튜터링이 효과적인 정도는 문제 해결 단계 내 또는 단계 후에 피드백을 제공하는 방식에 따라 달라진다고 했습니다. 그는 단계 기반 튜터링이 서브단계 튜터링만큼 효과적이라는 증거를 발견했으며, 따라서 단계 기반 튜터링이 최적의 곡물 크기라고 결론지었습니다.

피드백은 또한 매우 중요합니다. Shute(2008)에 따르면, 적응형 피드백은 제공되는 정보의 양(예: 확인 피드백, 힌트, 상세한 피드백), 피드백의 시기(예: 즉시, 지연된 피드백), 그리고 피드백의 목표(예: 즉각적인 다음 단계 안내, 학습 목표 달성의 진행 상황에 대한 안내) 등에 따라 달라질 수 있습니다. 적응형 기능(예: 개인화된 피드백)은 매크로 수준에서 학습 목표를 달성하기 위한 최적의 과제를 선택하거나, 마이크로 수준에서 현재 과제의 다양한 측면과 피드백의 수준을 조정하는 방식으로 제공될 수 있습니다(VanLehn et al., 2007).

6.3 진단 평가와 과정 분석

인지 진단 모델링(CDM)은 문제 해결과 관련된 인지 처리 요구 사항을 나타내는 특징으로 코드화된 문제 항목이나 과제에 대한 반응을 모델링하는 방법의 집합입니다. CDM을 사용하여 반응 데이터를 모델링하는 동기는 학습자의 기본적인 정보 처리 방식을 드러내기 위해서입니다. 이를 통해 학습자가 올바르게 또는 잘못 답한 항목의 패턴을 바탕으로 학습자가 무엇을 알고 무엇을 모르며, 그들이 가질 수 있는 오해를 추론할 수 있습니다. 인지

진단 모델링의 약속은 학습자의 문제 해결 특성을 진단 목적으로 드러내어, 학습과 피드백을 학습자에게 맞춤화할 수 있게 하는 것입니다. CDM의 동기와 접근법은 학생 모델링을 다루는 ITS 문헌의 동기와 접근법과 유사하며, 이 역시 맞춤화에 관심이 있지만, 최근까지 그들의 역사적 배경은 독립적이었습니다. CDM은 심리측정학의 한 분야이고(von Davier, 2010), ITS 학생 모델링은 완전히 인지 심리학의 학습 문헌에서 발전하였으며(Corbett & Anderson, 1994), 학생 모델링을 위해 '지식 추적'이라는 방법을 채택했습니다(Liu, Kell, et al., 2023).

인지 진단 모델링과 평가에 대한 연구는 오랜 역사와 방대한 문헌이 있습니다(Rupp et al., 2010). 최근에는 응답 시간과 같은 과정 데이터를 통합하여 개인의 학습 과정을 더 잘 이해하려는 노력이 진행되고 있습니다(Zhan et al., 2018). 또한, CDM과 ITS 학생 모델링 문헌을 연결하려는 시도도 있었습니다(Wang et al., 2018). 한 가지 접근 방식은 CDM을 Bayesian 지식 추적(BKT)에 사용하는 것으로, 이는 숨겨진 마르코프 모델(HMM)을 기반으로 ITS 문헌에서 사용되는 학생 모델링 방법입니다(Wang et al., 2018, 2020). Wang et al. (2018)은 BKT HMM과 CDM 프레임워크를 결합하여 여러 기술의 성장 추적을 가능하게 하고, HMM 기술 전이를 모델링하기 위해 공변수를 수용할 수 있도록 했습니다. 이 분야는 빠르게 성장하고 있으며, 점점 더 정확하고 해석 가능하며 실행 가능한 인지 진단을 통해 맞춤화가 강화될 것으로 기대됩니다(Wang et al., 2020).

6.4 피드백

지금까지의 논의는 피드백이 많은 교육 개입의 핵심 요소임을 시사합니다—형성 평가, 테스트 효과, 인간 및 기계 튜터링 등에서 중요한 역할을 합니다. 또한, 피드백이 교육적 성과에 미치는 영향에 대한 독립적인 연구 문헌도 존재합니다. 컴퓨터 기반 교육의 잠재력을 보여준 초기 연구에서 Azevedo와 Bernard(1995)는 컴퓨터 기반 교육에서 피드백이 즉각적인 성취도 사후 테스트에서 0.80, 지연된 사후 테스트에서 0.35의 효과 크기를 나타냈다고 보고했습니다. Hattie와 Timperley(2007)는 0.79의 효과 크기를 추정했으며, 최근 Wisniewski et al.(2020)은 더 엄격한 배제 규칙을 사용하여 0.48의 효과 크기를 추정했지만 상당한 이질성이 있었습니다. 피드백 효과는 인지 및 운동 기술에 비해 동기 및 행동 기술에서 더 컸습니다. 또한, 피드백은 제공되는 정보의 양에 따라 더 효과적이었으며, 실수의 원인과 그 해결 방법을 학생들이 이해할 수 있도록 도와주는 것이 가장 유익했습니다. 피드백의 타이밍도 중요한 요소로 밝혀졌습니다(Hattie & Timperley, 2007). 즉각적인 피드백은 종종 더 효과적이지만, 학습자가 복잡한 과제에 참여하는 경우 지연된 피드백이 더 효과적일 수 있습니다(예: Attali & van der Kleij, 2017; Fyfe et al., 2021; Hattie, 2009).

효과적인 피드백은 교육적 맥락, 과제의 성격, 학습자의 특성을 고려해야 한다고 Shute (2008)는 제안했습니다. 무엇이 가장 효과적인지는 상황에 따라 달라질 수 있습니다. Panadero와 Lipnevich (2022)는 다양한 상황에서 효과적일 수 있는 피드백의 통합적인 유형을 제시했습니다. 이 유형은 피드백을 내용(예: 확인, 설명), 기능(예: 학습 지원, 동기 부여, 숙달 지향성 증진), 제시 방법(예: 즉시성, 빈도, 학습자의 진전에 따른 적응성, 피드백을 전달하는 매체의 수), 출처(예: 교사, 동료, 자기, 컴퓨터)로 분류합니다.

정보성 피드백은 성취도뿐만 아니라 학습자의 참여도, 노력, 지속성, 만족도와 같은 동기 변수에도 영향을 미칩니다 (Narciss, 2004). Shute (2008)는 효과적인 교수 피드백이 여러 가지 특성을 가져야 한다고 주장했습니다. 피드백은 다음과 같은 특성을 가져야 합니다:

편향되지 않아야 함: 피드백은 공정하고 균형 잡혀야 합니다 (Kluger & DeNisi, 1996; Panadero, 2023). 학습자가 아니라 과제에 초점을 맞춰야 함: 피드백은 학습자가 아닌 과제에 초점을 맞추어야 합니다 (Fyfe et al., 2023). 구체적이고 명확해야 함: 피드백은 학습자의 오해를 해결하고 장기적인 학습을 유도할 수 있도록 구체적이고 명확한 형식으로 제시되어야 합니다 (Attali & van der Kleij, 2017; Moreno, 2004). 학습자가 과제를 시도한 후 제공되어야 함: 피드백은 학습자가 학습 과제를 시도한 후에 제공되어야 합니다 (Hattie & Gan, 2011). 지속적인 학습을 촉진해야 하며, 학습자의 현재 성과와 의도된 학습 결과 사이의 불일치를 줄여야 함: 피드백은 학습자의 성과와 목표 간의 차이를 줄이고 학습이 계속해서 이루어지도록 돕는 역할을 해야 합니다 (Leenknecht et al., 2019).

6.5 혁신적 평가 설계에 대한 시사점

효과적인 피드백은 학습자가 자신의 학습을 개선할 방법과 그 개선을 도울 수 있는 자원을 어떻게 사용할지 안내하는 도구입니다 (Hattie & Timperley, 2007). 효과적인 피드백의 특성을 이해하는 것은 인간 간 피드백의 이점을 훨씬 더 큰 규모로 제공할 수 있는 디지털 학습 및 평가 시스템의 설계에 도움이 될 수 있습니다. 이러한 지식은 혁신적인 평가 시스템 내에서 피드백 기능을 구조화하는 데 사용되어 모든 학습자가 특정 학습 목표에 맞는 다양한 유형의 피드백을 받을 수 있도록 보장할 수 있습니다.

미래의 디지털 평가 설계는 각 학습자의 강점과 약점에 맞춘 개인화된 피드백을 제공하는 방법을 고려해야 하며 (Panadero & Lipnevich, 2022), 이를 통해 보다 효과적인 학습 경험을 이끌어낼 수 있습니다. 모든 학생이 공평한 학습 경험에 접근하고 참여할 수 있는 기회를 가질 수 있도록 하기 위해, 디지털 평가는 명확하고 접근 가능하며 다양한 학생의 요구를 수용하는 피드백을 제공해야 합니다. 잘 설계된 개인화된 피드백은 또한 매우 동기 부여가 될 수 있으며, 학습을 향상시킬 수 있는 정보만 제공하는 것이 아니라 학습 과제와 관련된 더 큰 흥미와 가치를 전달할 수 있습니다 (Narciss et al., 2014). 평가 설계에 따라 피드백은 교사, 동료 또는 시뮬레이션된 에이전트와의 대화형 상호작용을 통해 학습자를 참여시킬 수 있으며, 이는 학습 경험을 더 상호작용적이고 협력적이며 흥미롭게 만들 수 있습니다.

디지털 평가에서 학습자를 위한 피드백을 이해하고 포함시키는 것은 여러 면에서 평가의 혁신과 개선으로 이어질 수 있습니다. 디지털 학습 플랫폼 내에서 피드백을 찾는 행동은 스스로 조절하는 학습(self-regulated learning)과 관련된 행동을 나타내는 지표가 될 수 있습니다. 예를 들어, 이러한 행동과 관련된 클릭스트림 데이터를 분석하면 학생들이 학습 과정을 어떻게 관리하고, 지도나 피드백을 어떻게 구하며, 이를 바탕으로 전략을 어떻게 조정하는지에 대한 통찰을 제공할 수 있습니다 (예: Aguilar et al., 2021; Bernacki, 2018; Ober et al., 2023; Tenison & Sparks, 2023). 이 이해는 자가 조절 학습을 촉진하고 학생들이 더 효과적인 학습 습관을 개발하도록 유도하는 평가 설계를 돕는 데 활용될 수 있습니다. 학습자에 대한 지원은 개별 학생이 피드백에 얼마나 반응하는지, 피드백에 따른 행동, 그리고

개선이 필요한 특정 영역에 맞춰 맞춤화될 수 있습니다. 다중모드 데이터 소스를 활용하면 학생들의 학습 행동에 대한 더 포괄적인 시각을 제공하고, 보다 효과적이고 개인화된 개입을 가능하게 합니다 (Lehman et al., 2018; Sparks et al., 2024; Zapata-Rivera, Lehman, & Sparks, 2020). 이러한 발전은 전체 학습 경험을 향상시키고 디지털 학습 환경에서 더 나은 교육적 결과를 이끌어낼 잠재력을 지니고 있습니다.

6.6 학습 원리

시험 응시자에게 피드백을 제공하는 기반은 학습 원칙에 뿌리를 두어야 합니다. 이전 섹션에서는 피드백 생성 및 전달을 지원하는 학습 원칙을 다루었지만, 지난 한 세기 동안 발전해 온 더 넓은 학습 원칙들을 고려하는 것도 유용합니다. 이 문헌은 방대하지만, 미래의 평가 맥락에서 특히 적합할 수 있는 몇 가지 유용한 종합이 존재합니다. Thorndike는 효과(강화), 연습(실습), 준비성(준비성)의 세 가지 학습 법칙을 제안했으며, 이는 여전히 유효합니다. 미국심리학회(APA, 2018)는 PreK–12 교육을 위한 심리학 원칙 중 상위 20가지를 제시했으며, 여기에는 사고 및 학습, 동기 부여, 사회적·정서적 맥락, 교실 관리, 학생 진척도 평가와 관련된 원칙이 포함됩니다. 카네기 멜론 대학교의 Eberly Center(2024)는 효과적인 학습을 위한 일곱 가지 원칙을 제시했으며, 그 중 일부는 학습에 도움이 되거나 방해가 될 수 있는 이전 지식, 지식 조직이 학습에 미치는 영향, 동기가 학습 행동을 지배한다는 점, 기술 구성 요소를 결합하고 목표 지향적인 연습과 피드백이 중요하다는 점, 사회적·정서적 요소와 지적 요소가 모두 중요하다는 점, 그리고 자기 모니터링과 조정이 자기 주도적 학습자가 되는 데 중요하다는 점을 강조합니다. Schwartz et al. (2016)은 교육자들이 사용할 수 있도록 설계된 26개의 학습 원칙에 대한 근거 기반 요약을 제공했습니다.

Bjork와 Bjork(2011)는 '바람직한 어려움(desirable difficulties)'이라는 개념을 중심으로 학습 원칙들을 강조했습니다. 바람직한 어려움은 학습에 어려움을 주지만, 그 결과 더 오래 가고 유연한 학습을 유도하는 학습 조건을 의미합니다. 여기에는 연습 조건을 다양하게 설정하기, 시험 전에는 벼락치기보다는 학습 세션과 연습 세션을 간격을 두고 배치하기, 전체적인 학습의 일부로서 학습해야 할 과제들을 차례대로(차단하지 않고) 가르치는 것, 그리고 생성 효과 및 관련된 시험 효과가 포함됩니다(앞서 다룬 내용). 각 조건에서 쉬운 조건은 단기적으로 성과를 올릴 수 있지만, 어려운 조건은 장기적인 성과와 습득한 지식을 더 유연하게 활용할 수 있게 하므로, 바로 이러한 조건들이 바람직한 어려움입니다.

국립연구위원회(National Research Council, 2000)와 국립과학기술학회(National Academies of Science, Engineering and Medicine, 2018)는 다양한 학문 분야에서 학교 및 직장 환경을 아우르는 학습 원칙을 요약한 **"How People Learn"**과 **"How People Learn II"**라는 두 권의 포괄적인 시리즈를 발행했습니다. 이 시리즈는 문화, 학습 유형, 지식과 추론, 동기, 학교 학습, 기술, 생애 전반에 걸친 학습 등 다양한 주제를 다루며 결론을 도출하고 있습니다. 또한, 학습 맥락과 기술이 학습에 미치는 중요성에 관한 향후 연구에 대한 권장 사항도 제시하고 있습니다. 이 두 권과 그 외의 자료에서 다룬 학습 원칙들은 평가 맥락에서 적용될 수 있는 피드백 개발을 안내하는 데 사용할 수 있습니다.

6.7 결론: 피드백

이 섹션에서 다룬 문제는 테스트가 종종 시간, 노력, 비용 면에서 수험자에게 많은 요구를 하지만, 그에 비해 교육적 가치는 거의 제공하지 않는다는 점입니다. 질문은 시험이 수험자에게 어떤 가치를 제공할 수 있을까요? 이 섹션에서는 평가와 테스트가 수험자나 평가 대상에게 유용한 정보를 제공하거나 기술 습득을 돕는 여러 방법을 살펴보았습니다. 또한, 시험이 제공할 수 있는 정보의 가치나 지원에 대한 근거 기반 추정치를 제공했습니다. 우리는 시험 점수와 종종 제공되는 규범적 기준 외의 정보를 중점적으로 다루었으며, 규범적 및 해석적 정보가 여전히 중요한 가치를 제공한다는 점도 언급했습니다.

형성 평가는 테스트를 교육 과정의 중요한 부분으로 사용하는 것으로, 매우 다양한 방식으로 구현되지만, 학습에 상당히 긍정적인 영향을 미친다는 것이 입증되었습니다. 테스트 효과, 또는 테스트 연습은 학습자의 공부 시간을 일부 테스트 시간으로 대체하는 것인데, 이는 학습 결과에 강한 긍정적인 효과를 미친다고 보여졌습니다. 인간 튜터링은 가장 강력한 교육적 개입 중 하나로 밝혀졌으며, 컴퓨터 기반 지능형 튜터링 또는 적응형 교육도 마찬가지로 강력한 개입으로 확인되었습니다. 인간 또는 기계 튜터링이 강력한 이유는 완전히 이해되지 않았지만, 피드백 제공, 유도된 프롬프트, 상호작용 촉진 및 건설적인 행동을 장려하는 것이 중요한 요소라는 증거가 있습니다. 튜터링은 또한 학습자의 인지 진단을 수행하며, 테스트도 유사한 역할을 합니다. 점점 더 정교해지는 인지 진단 모델링은 AI 기술의 발전을 활용하고, 테스트 응답자의 더 많은 과정 행동을 학습자 모델에 통합하여 학습자에게 맞춤형 지도를 제공하는 데 유용한 지원을 약속합니다.

피드백 역시 학습을 개선하는 강력한 수단으로 개인화가 중요한 역할을 한다는 것이 밝혀졌습니다. 어떤 종류의 피드백이 가장 효과적인지에 대해서는 이미 많은 것이 알려져 있으며, 생성적 AI를 활용하여 학습자와 학생들에게 유용한 피드백을 제공하는 것은 유망한 새로운 방향입니다. 마지막으로, 학습 과정 자체에 대해 우리는 몇 십 년 전보다 훨씬 더 많은 것을 알고 있으며, 이를 통해 향상된 학습 결과를 도출할 수 있는 방법도 많이 알게 되었습니다. 피드백, 지도, 학습 지침을 수립할 때 증거 기반 학습 원칙을 따르는 것이 평가의 가치를 상당히 향상시킬 것입니다.

따라서 평가는 학습자가 자신의 기술 수준에 대한 정보를 교사나 정책 입안자에게 제공하고, 동시에 그들 또한 학습 여정에서 다음에 무엇을 해야 할지에 대한 지침을 받아 현재의 기술 수준과 학습 목표 사이의 격차를 좁히고, 기술을 향상시키며, 자율성, 능숙함, 소속감을 개발하는 두 가지 방향으로 이루어질 수 있습니다. 적절하게 설계되고 개인화된 피드백 제공은 교육의 공평성을 위한 목표를 달성하고 모든 학습자의 학습과 성과를 촉진하는 데 기여할 수 있습니다.

7 요약 및 결론

이 논문의 목적은 현재 평가 분야의 상태를 검토하고, 평가의 미래가 어떻게 될지에 대해 추측하는 것이었습니다. 우리는 검토를 바탕으로 유망한 평가 연구 방향을 고려했습니다. 평가의 미래는 주로 교육과 직업의 미래, 그리고 사회가 미래에 필요로 할 기술들에 관한 문제이며, 따라서 우리는 먼저 미래에 가장 중요한 기술이 무엇일지 고려했습니다. 미래 기술에 대한 우리의 분석은 트렌드 분석, 고용주 설문조사, 기술의 영향 분석, 전문가 의견, 그리고 17개 중상위 소득 국가의 17,000명 성인을 대상으로 한 ETS(2023a) 조사 결과를 바탕으로 했습니다. 그다음으로 우리는 이러한 기술을 측정하는 혁신적이고 유망한 접근 방식을 고려했으며, 특히 측정하기 어려운 기술을 측정하는 방법에 집중했습니다. 그 후, 우리는 시험 운영을 고려했으며, 이는 관리 측면에서부터 문항 개발, 개인화, 보안, 채점 및 평가까지 포함됩니다. 우리는 또한 AI와 기술이 이러한 운영을 어떻게 향상시킬 수 있는지에 대해 강조했습니다. 마지막으로, 우리는 학습 과학의 관점에서, 그리고 시험 응시자 및 기타 이해관계자들의 요구에 기반하여 시험 응시자에게 제공되는 피드백에 대해 고려했습니다.

우리는 연구 결과를 바탕으로 여러 가지 중요한 결론을 도출했습니다. 첫째, 특히 AI를 중심으로 한 기술의 발전은 평가의 모든 측면에 깊은 영향을 미칠 것이며, 우리는 그 영향을 지금 막 이해하기 시작했다고 할 수 있습니다. 이러한 변화는 측정할 기술이 무엇일지부터, 그것들을 어떻게 측정할 것인지, 시험 결과를 응시자와 이해관계자에게 어떻게 보고할 것인지, 그리고 결과를 받은 사람들이 그 결과를 어떻게 활용할 것인지까지 모두 포함됩니다.

두 번째로, 소프트 스킬, 지속 가능한 스킬, 복잡한 스킬의 핵심적인 세트가 미래에 점점 더 중요해질 가능성이 큽니다. 특히 교육 성취도와 직장 기술의 평가 역사에서, 그동안 주로 커리큘럼과 기술적 스킬에 초점이 맞춰졌습니다. 이러한 기술들의 평가는 계속 중요할 것이며, 특히 이러한 기술들의 변화와 성장에 대한 평가가 중요합니다. 하지만 소프트 스킬이 학교, 직장, 그리고 삶에서 성공을 위해 기술적 스킬만큼, 아니 더 중요할 수 있다는 새로운 인식이 생겼습니다. 사회적 스킬—팀워크, 협업, 커뮤니케이션—은 직업 트렌드에 따라 점점 더 중요해질 것입니다. 적응력은 AI와 기술 변화가 직장 내 업무 요구를 바꾸어 가는 상황에서 더욱 중요해질 것입니다. 이는 평생 지속적인 학습의 중요성을 부각시키며, 재정적 안정뿐만 아니라 개인의 만족과 웰빙에도 영향을 미칩니다. 창의력과 비판적 사고는 점점 더 중요해질 것입니다. 왜냐하면 이는 인간이 컴퓨터보다 우위에 있는 기술들로, 이는 일정 기간 동안 계속 유지될 가능성이 높으며, AI에 의해 대체되기보다는 보강될 것입니다.

이처럼 기술의 중요성이 증가함에 따라, 기술 개발을 평가하고 인정하는 시스템이 마련될 것입니다. 전 세계의 많은 응답자들이 비학위 자격증이 기술을 보여주는 중요한 방법이 될 것이며, 미래에는 특정 기술을 증명하는 그런 자격증이 대학 학위보다 더 중요해질 것이라고 믿고 있습니다. 이러한 자격증은 대학에서 발행될 수도 있지만, 회사나 표준화된 시험 또는 학습 평가 기관에서 발행되는 경우에도 동등하게 가치 있게 여겨질 것입니다. 기술 습득을

인증하는 마이크로 자격증 및 기타 인증을 얻기 위해 평가에 의존하게 되면, 이러한 인증의 보안 문제가 더욱 중요해질 것입니다.

네 번째 결론은 미래에 점점 더 중요해질 기술들에 대해 현재 우리가 가진 평가 방법들이 충분하지 않다는 것입니다. 많은 기술에 대한 평가 도구의 품질에 대해 회의적인 시각이 존재하는데, 이는 종종 인상이나 자기 보고서, 기타 주관적인 방법들에 의존하기 때문입니다. 이는 현재 측정하기 어려운 기술들에 대한 엄격하고 심리측정적으로 신뢰할 수 있는 평가 도구를 개발할 수 있는 거대한 기회를 제공합니다.

마지막으로, 평가에 대한 태도는 매우 긍정적입니다. 평가가 시험 응시자에게 새로운 기술을 습득하도록 동기를 부여하고, 기회를 추구하고 경력을 발전시키는 데 있어 자신감과 준비성을 느끼게 해 준다고 여겨집니다. 이는 AI 기반 변화가 직장에 미치는 영향으로 점점 더 중요해질 것입니다. 많은 사람들은 평가가 자존감을 높이고 경력 만족도를 증진시키며, 다양한 배경을 가진 사람들이 기회를 공평하게 제공받도록 기술 격차를 해소하는 역할을 한다고 보고 있습니다. 평가가 중요한 역할을 하려면, 응시자가 평가를 통해 자신에 대한 피드백과 통찰을 얻을 수 있는 방식에 달려 있습니다. 시험 응시자에게 개인화된, 유용하고 실행 가능한 피드백을 제공하는 것은 미래 평가의 중요한 목표이며 실현 가능한 목표입니다.

7.1 한계

평가의 미래를 예측하는 데에는 한계가 있습니다. 이는 일반적으로 미래를 예측하는 데 한계가 있다는 점과 같습니다. 사람들은 미래를 예측하는 데 그다지 정확하지 않습니다 (Grossmann et al., 2023; Rees, 2021). 그러나 평가의 일부를 설계하는 데 부분적인 책임이 있는 조직으로서, ETS는 단순히 미래를 예측하는 사람들보다 장점이 있을 수 있습니다. Grossmann et al. (2023)은 더 나은 예측자는 "예측 분야에 대한 과학적 전문성이 있고, 학제 간 접근을 하며, 더 간단한 모델을 사용하고, 이전 데이터를 바탕으로 예측을 한다"고 제안했습니다. 보고서 작성 팀과 검토자들은 다양한 관점에서 평가 전문성을 갖추고 있으며, 우리는 ETS 인간 발전 연구(ETS, 2023a)에 참여한 인터뷰 대상자의 전문성에 의존했습니다. ETS 인간 발전 연구의 외부 전문가들과 보고서 작성 팀을 통해 우리는 학제 간 관점에서 작업을 접근했고, 폭넓은 문헌을 검토하면서 이전 데이터를 바탕으로 했습니다. Rees (2021)은 개별 편향을 극복하기 위해 군중 소싱 전략을 제안했으며, 우리는 17개국에서 다양한 배경을 가진 17,000명 이상의 응답자 데이터를 제공하는 ETS 인간 발전 연구가 그 역할을 한다고 주장할 수 있습니다. 그럼에도 불구하고 우리는 체계적인 예측 방법론을 사용하지 않았으며, 따라서 평가의 미래에 대한 우리의 예측은 신중하게 해석되어야 합니다.

또 다른 한계점은 우리가 모든 평가 분야에 균등한 관심을 기울이지 않았다는 점입니다. 대신 기술, AI, 학습 과학 발전, 그리고 평가를 응시자에게 더 유용하게 만들 수 있는 기회와 같은 가능성과 현재 상태 간의 격차로 인해 가장 큰 변화를 겪을 분야에 집중했습니다. "미래의 기술" 섹션에서 우리는 기술-driven 변화로 인해 중요성이 커질 것으로 예상되는 기술에 초점을 맞추었습니다. National Research Council (2012)에서 다룬 기초적인 읽기, 수학, 기타

교과목의 숙달과 적용을 촉진하는 기술과 지식은 여전히 중요하게 유지될 것입니다. 우리가 다룬 새로운 측정 방법, 운영, 피드백에 관한 주제는 기초 문해력 및 기타 교과목뿐만 아니라 두 번째 섹션의 핵심이었던 새로운 내구성 있는 기술에도 적용됩니다. 그러나 우리는 K-12 평가에서 중요한 두 가지 분야인 교실 내 평가와 책임에 대해서는 집중하지 않았습니다. 이 보고서에서 다룬 기술과 전통적인 학문적 기술 및 읽기, 수학, 과학과 같은 내용 간의 관계를 더 깊이 탐구하고, 교실 내 평가와 책임에 중점을 두는 동반 보고서가 있을 수 있음을 상상할 수 있습니다. 이러한 보고서는 아마도 국제적으로 확장될 수 있습니다.

7.2 미래 방향

우리는 이 논문의 주요 섹션과 일치하는 몇 가지 주요 연구 문제와 방향을 제안하고자 합니다. 첫째, 기술의 변화가 요구하는 기술을 모니터링하는 것이 중요합니다. 노동 시장에서 요구되는 기술은 교육 기준과 교육 과정에 영향을 미치기 때문에 이러한 변화를 예측하는 것이 유용합니다. 둘째, 평가 내 학습 특성화 및 협력적이고 다중 모드 접근을 포함한 새로운 혁신적 접근 방식을 탐구하는 풍부한 평가 방법이 연구의 주목을 받을 가능성이 높습니다. OECD(2022a)의 복합 기술에 대한 혁신적 평가 고려 사항은 추구할 가치가 있는 제안으로 보입니다. 셋째, 시험 운영의 여러 측면—문항 개발, 개인화, 채점, 보안, 보고—은 이미 기술과 AI의 빠른 발전에 영향을 받고 있으며, 이러한 운영의 변화 속도는 느려지지 않을 것입니다. 마지막으로, 지난 10년 동안 발표된 수많은 평가 미래에 관한 논문에서 시험 응시자에게 유용하고 실행 가능한 피드백을 제공하여 그들이 어디에 있는지, 어떻게 개선할 수 있는지를 통찰할 수 있게 해 주는 방법에 대한 진전을 예고하거나 요청한 바 있습니다. 우리는 이러한 요청을 지지합니다.

ETS 연구소는 네 가지 연구 분야를 통해 이러한 방향에 대응하고 있습니다. 이 연구들은 평가 개인화, 혁신적이고 상호작용적인 디지털 평가 창출을 위한 설계 원칙 수립, 자동화된 콘텐츠 생성 및 채점을 포함한 책임감 있고 윤리적인 AI 응용 프로그램의 표준 개발, 그리고 불평등을 해소하는 차세대 교육 시스템을 구상하여 정책과 실천에 영향을 미치는 것에 중점을 둡니다. 이곳에서 설명한 연구와 ETS 연구소의 연구 분야를 통해 우리는 전통적인 성취 및 능력 측정의 역할을 유지하면서도 인간 학습에 더 나은 서비스를 제공할 수 있도록 평가를 재구성할 수 있는 위치에 있을 것입니다. 이는 2013년 Gordon 위원회에서 발표한 '미래 교육 평가'에 대한 논문 모음에서 제시된 평가의 비전에 한 걸음 더 가까워질 것입니다.

마지막으로, 이러한 비전을 실현할 수 있는 교육 및 기술 평가의 발전을 촉진하기 위해, 우리는 상당한 연구 투자가 필요하다고 주장합니다. 전 세계 교육 지출은 매년 5조 달러 이상으로, 이는 전 세계 국내총생산(GDP)의 약 6%에 해당합니다 (World Economic Forum, 2022). 그러나 그 중에서 평가와 관련된 부분은 매우 적으며, 이는 인간 학습을 지원하고 교육의 진전을 모니터링하는 데 필요합니다. World Economic Forum(2021)의 '직장에서 필요한 기술의 글로벌 분류'는 기술 기반 노동 시장에 대한 비전을 제시합니다. 동반 보고서인 World Economic Forum(2023)의 'Education 4.0' 프레임워크는 글로벌 시민성, 혁신과 창의성, 기술 능력, 대인 관계 능력 등의 내용을 기술로 제시하며, 이는 다음 세대가 미래의 직장과 사회를 준비하는 데 중요하다고 강조합니다. 이 세 가지 기술은 본 보고서의 '미래의 기술' 섹션에서

현재와 미래의 노동 시장 수요 분석을 바탕으로 확인한 기술들과 잘 일치합니다. Education 4.0은 또한 개인화되고 자기 주도적인, 접근 가능하고 포용적인, 문제 기반의 협력적인, 평생 학습을 강조하는 학습 경험의 중요한 발전을 제시하는데, 이는 본 보고서에서 제시한 주제와 잘 일치합니다. 평가의 발전은 집중과 투자를 통해 달성할 수 있으며, 이는 Gordon 위원회(2013)와 World Economic Forum(2021, 2023)의 보고서에서 제시된 비전을 실현하는 데 중요한 역할을 할 것입니다.

8 감사의 말

Kadriye Ercikan, Ida Lawrence, Michelle Froah, Sarah Rhame, Christine Betaneli, Kateryna Komarova, Charlotte Kirkby, Matthew Johnson이 이 보고서에 기여했습니다.

8.1 주석

1 ETS 인간 진보 연구(ETS Human Progress Study, ETS, 2023a)는 미래에 대한 9명의 세계적 사상 지도자들과의 심층 인터뷰와 해리스 여론조사(Harris Poll)와 협력하여 실시한 설문조사로 구성되어 있습니다. 이 설문조사는 2023년 9월 18일부터 27일까지 17개 고소득 및 중소득 국가(국가당 최소 1,000명)의 18세 이상 17,143명의 응답자를 대상으로 평가의 미래 및 기타 사회 문제와 사회적 결과와 관련된 다양한 주제에 대한 그들의 견해에 대해 실시되었습니다. 우리는 이 논문 전반에 걸쳐 사상 지도자들을 인용하고 연구의 설문조사 결과를 제시합니다. 데이터는 전체 인구의 대표성을 보장하기 위해 가중치가 부여되었습니다. 그러나 데이터는 전체 국가 인구로 일반화되지 않을 수 있으며, 결과는 인구의 의견이 아닌 다양한 표본의 의견으로 보아야 합니다. 국가는 세계은행 정의에 따라 고소득 또는 중소득 국가로 언급되었습니다.

2 타당도는 일반적으로 검사 및 심리학 내외의 다른 맥락에서 관련되지만 다른 해석을 가지고 있습니다. 어떤 관점에서는 타당도는 해석이 아니라 측정 도구의 속성입니다(Hood, 1998 참조). 교육 및 심리 측정 내에서의 다양한 관점은 Lissitz(2009)를 참조하십시오.

3 고부담(high-stakes)과 저부담(low-stakes)은 또한 연속체로 이해될 수 있습니다. 의사 결정에서 거의 비중을 차지하지 않는 고부담 시험은 유일한 결정 요인으로 작용하는 고부담 시험과 같지 않습니다. 그리고 보다 일반적으로, 정의에 따른 "중요한" 그리고 "직접적인" 결과는 각각 높음에서 사소함까지, 직접에서 간접까지 연속적으로 다양할 수 있습니다. Tannenbaum과 Kane(2019)은 추가적인 고려사항을 제시했습니다.

4 시험 공정성 문제는 International Test Commission(2001, 2013, 2017), International Test Commission and Association of Test Publishers(2022), Society for Industrial and Organizational Psychology(2018)에서도 고려됩니다.

5 무시험 평가(즉, 자연 발생적인 행동을 측정하는 평가)는 매우 관심이 집중되는 주제가 되었습니다. 2024년 유럽연합 AI법(https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf)은 "직장 및 교육 기관 분야에서 AI 시스템을 사용하여 자연인의 감정을 추론하는 것"을 금지하며(108쪽), 안전 및 의료 예외를 두고 있으며, "공공 또는 민간

주체가 자연인의 사회적 점수를 매기는 데 AI 시스템을 사용하면 차별적인 결과와 특정 집단의 배제로 이어질 수 있다"고 경고합니다(26쪽).

6 이 기사의 범위를 벗어나지만, 기술의 지리학(Moro et al., 2021)과 같은 미래 기술 우선순위를 결정하는 데 있어 중요한 추가 고려사항이 있습니다.

7 자세한 내용은 주석 6을 참조하십시오.

8 여기서 우리는 통계학에서의 기술 용어가 아닌 편향이라는 용어의 원래 의미를 사용합니다.

9 자동 에세이 채점 문헌에서 정확도 또는 정확히 일치하는 정도는 일반적으로 두 평가자 간 또는 기계와 인간 점수 간의 정확한 일치로 측정됩니다. 정확도는 또한 카파(kappa), (선형) 가중 카파(weighted kappa) (Cohen, 1968), 이차 가중 카파(quadratic weighted kappa) (선형 패널티를 넘어 불일치에 패널티를 부과함) 측정에서와 같이 무작위 기회의 기준선으로 정규화될 수 있습니다. 일반적인 접근 방식은 인간-인간 점수 일치도의 저하를 고려하는 것입니다. 이러한 측정을 사용하여 Williamson 등(2012)은 많은 유형의 에세이에 대해 인간 간 합의로부터 최소한의 저하만 있었다고 보고했으며, 실제로 "자동화된 인간 간 합의가 인간-인간 합의보다 높은 것을 관찰하는 것은 비교적 일반적"이라고 했습니다(8쪽). 트랜스포머 기반 접근법을 사용한 최근 연구(Ormerod et al., 2021)에서는 인간 수준 이상의 성능을 보고했습니다.

10 이 기사에서는 효과 크기에 대해 언급하는데, 이는 결과에 대한 조작(또는 관계)의 강도를 나타내는 지표입니다. Cohen(1992)이 제안한 고전적인 경험 법칙에 따르면 작은, 중간, 큰 효과 크기는 각각 .20, .50, .80보다 큰 효과 크기 값에 해당합니다.

11 VanLehn(2011)은 Bloom(1984)의 종료 조건(다음 수업으로 넘어가는 데 필요한 숙달 수준)이 다양했다고 주장했으며, 따라서 Bloom은 실제로 숙달의 효과에 대한 증거를 제시했습니다.

9 감사의 말씀

Kadriye Ercikan, Ida Lawrence, Michelle Froah, Sarah Rhame, Christine Betaneli, Kateryna Komarova, Charlotte Kirkby, Matthew Johnson이 이 보고서에 기여하였습니다.

9.1 주석

1. ETS Human Progress Study (ETS, 2023a)는 세계적인 미래 분야의 사상가 9명과의 심층 인터뷰 및 Harris Poll과 협력하여 진행된 설문조사입니다. 이 설문조사는 2023년 9월 18일부터 27일까지 17개 고소득 및 중소득 국가에서 18세 이상 17,143명의 응답자를 대상으로 진행되었으며, 응답자들은 평가의 미래와 기타 사회적 문제 및 사회적 결과와 관련된 다양한 주제에 대한 의견을 제공했습니다. 본 논문에서는 사상가들의 의견을 인용하고, 연구의 설문 결과를 제시합니다. 데이터는 전체 인구를 대표할 수 있도록 가중치가 부여되었습니다. 그러나 이 데이터는 전체 국가 인구를 일반화할 수 없으며, 결과는 다양한 샘플의 의견으로 해석되어야 합니다. 국가들은 세계은행의 정의에 따라 고소득 국가 또는 중소득 국가로 분류되었습니다.
2. 타당도는 테스트와 심리학 분야 내외에서 다르게 해석될 수 있습니다. 일부 관점에서는 타당도가 해석이 아니라 측정 도구의 특성으로 간주됩니다 (Hood, 1998 참조). 교육 및 심리학적 측정에서의 다양한 관점에 대해서는 Lissitz (2009)을 참조하세요.
3. 고위험과 저위험은 연속체로 이해될 수 있습니다: 결정에 미치는 영향이 적은 고위험 테스트는 유일한 결정 요인으로 작용하는 고위험 테스트와 같지 않으며, 일반적으로 "중요한"과 "직접적인" 결과는 각각 고위험에서 미미한 위험, 직접적인 결과에서 간접적인 결과로 연속적으로 변할 수 있습니다. Tannenbaum과 Kane (2019)은 추가적인 고려사항을 제시했습니다.
4. 테스트 공정성 문제는 또한 International Test Commission (2001, 2013, 2017), International Test Commission과 Association of Test Publishers (2022), Society for Industrial and Organizational Psychology (2018)에서 다루어졌습니다.
5. 시험 없는 평가(즉, 자연적으로 발생하는 행동을 측정하는 평가)는 매우 논란이 많은 주제가 되었습니다. 2024년 유럽연합 AI 법률(https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf)은 "직장과 교육 기관에서 자연인 감정을 추론하는 AI 시스템의 사용"을 금지하며 (p. 108), "공공 또는 민간 주체가 수행하는 자연인에 대한 사회적 평가가 차별적인 결과와 특정 그룹의 배제를 초래할 수 있다"고 경고합니다 (p. 26).

6. 이 기사의 범위를 벗어나지만, 미래의 기술 우선순위를 결정하는 데 중요한 추가 고려사항이 있으며, 그 중 하나가 기술의 지리적 분포입니다 (Moro et al., 2021).

7. 추가 정보를 원하시면 주석 6을 참조하세요.

8. 여기서 '편향'이라는 용어는 통계학에서의 기술적 의미가 아니라 원래의 의미로 사용됩니다.

9. 자동화된 에세이 점수화 문헌에서는 정확도 또는 정확한 일치를 두 평가자 또는 기계와 사람의 점수 간의 정확한 일치로 일반적으로 측정합니다. 정확도는 또한 무작위 기회라는 기준으로 정규화될 수 있으며, 예를 들어 카파, (선형) 가중 카파 (Cohen, 1968), 그리고 선형 처벌을 넘는 차이를 처벌하는 이차 가중 카파 측정이 있습니다. 일반적인 접근법은 사람-사람 점수 일치에서의 퇴화를 고려하는 것입니다. 이러한 측정 방법을 사용한 Williamson et al. (2012)은 많은 유형의 에세이에서 사람 간 점수 일치에서 퇴화가 거의 없음을 보고했으며, 실제로 "자동화된-사람 점수 일치가 사람-사람 점수 일치보다 높은 경우가 상대적으로 흔하다" (p. 8) 고 밝혔습니다. 최신 연구들에서는 트랜스포머 기반 접근법을 사용하여 사람 수준을 넘는 성과를 보고했습니다 (Ormerod et al., 2021).

10. 이 기사에서는 효과 크기를 언급하며, 이는 결과에 대한 조작의 강도를 나타냅니다. Cohen (1992)이 제안한 고전적인 규칙은 작은, 중간, 큰 효과 크기가 각각 .20, .50, .80을 초과하는 값과 일치한다고 합니다.

11. VanLehn (2011)은 Bloom(1984)의 종료 조건(다음 레슨으로 넘어가기 위해 필요한 숙련도가 다를 수 있음)을 주장하며, Bloom은 사실 숙련도의 효과에 대한 증거를 제시했다고 합니다.

References

AAMC. (n.d.). AAMC PREview Professional Readiness Exam. https://students-residents.aamc.org/aamc-preview/aamc-previewprofessional-readiness-exam

Abrams, Z. (2024). Addressing equity and ethics in artificial intelligence. Monitor on Psychology, 55(3), 24–29. https://www.apa.org/monitor/2024/04/addressing-equity-ethics-artificial-intelligence

Abyaa, A., Khalidi Idrissi, M., & Bennani, S. (2019). Learner modelling: systematic review of the literature from the last 5 years. Educational Technology Research and Development, 67, 1105–1143.

Acar, S. (2023). Creativity assessment, research, and practice in the age of artificial intelligence. Creativity Research Journal, 1–7. Advance online publication. https://doi.org/10.1080/10400419.2023.2271749

Acuity Insights. (n.d.). What is Casper? https://acuityinsights.app/casper/

Acuity Insights. (2023). Casper technical manual. https://acuityinsights.com/casper-technical-manual/

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. Review of Educational Research, 87(3), 659–701. https://doi.org/10.3102/0034654316689306

Agrawal, A., Gans, J., & Goldfarb, A. (2022). Power and prediction: The disruptive economics of artificial intelligence. Harvard Business Review Press.

Aguilar, S. J., StuartA.Karabenick, S. A., StephanieD. Teasley, S.D.,Clare Baek,C. (2021).Associations between learning analytics dashboard exposure and motivation and self-regulated learning. Computers & Education, 162, Article 104085, https://doi.org/10.1016/j.compedu.2020.104085

Ahn, T., Arcidiacono, P., Hopson, A., & Thomas, J. R. (2019). Equilibrium grade inflation with implications for female interest in STEM majors (Working Paper 26556). National Bureau of Economic Research. https://doi.org/10.3386/w26556

Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. The Quarterly Journal of Economics, 134(3), 1121–1162. https://doi.org/10.1093/qje/qjz006

Ali, U. S., & van Rijn, P. W. (2016). An evaluation of different statistical targets for assembling parallel forms in item response theory. Applied Psychological Measurement, 40(3), 163–179. https://doi.org/10.1177/0146621615613308

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). Standards for educational and psychological testing. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). Standards for educational and psychological testing. American Psychological Association.

American Psychological Association. (2018). Top 20 principles from psychology for preK-12 teaching and learning: Coalition for psychology in schools and education. https://www.apa.org/ed/schools/teaching-learning/top-twenty-principles.pdf

Association of Test Publishers. (2022). Guidelines for technology-based assessment. https://www.testpublishers.org/assets/TBA%20Guidelines%203-14-2022%20draft%20numbered.pdf

Attali, Y.,&van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. Computers & Education, 110, 154–169. https://doi.org/10.1016/j.compedu.2017.03.012

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. Frontiers in Artificial Intelligence, 5, Article 903077. https://doi.org/10.3389/frai.2022.903077

Autor, D. H., Levy, F.,&Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. The Quarterly Journal of Economics, 118(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Autor, D., Chin, C., Salomons, A., & Seegmiller, B. (2024). New frontiers: The origins and content of new work, 1940–2018. The Quarterly Journal of Economics. Advance online publication. https://doi.org/10.1093/qje/qjae008

Azevedo, R., & Bernard, R. M. (1995). The effects of computer-presented feedback on learning from computer-based instruction: A meta-analysis. Journal of Educational Computing Research, 13(2), 111–127. https://doi.org/10.2190/9LMD-3U28-3A0G-FTQT

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–17. https://doi.org/10.5281/zenodo.3554657

Bailey, T., Jeong, D.W.,&Cho, S.W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. Economics of Education Review, 29(2), 255–270. https://doi.org/10.1016/j.econedurev.2009.09.002

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. Review of Educational Research, 61(2), 213–238. https://doi.org/10.3102/00346543061002213

Bauer, M. S., Damschroder, L., Hagedorn, H., Smith, J., & Kilbourne, A. M. (2015). An introduction to implementation science for the non-specialist. BMC Psychology, 3(32), 1–12. https://doi.org/10.1186/s40359-015-0089-9

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). A feasibility study of on-the-fly item generation in adaptive testing (Research Report No. RR-02-03). ETS. https://doi.org/10.1002/j.2333-8504.2002.tb01890.x

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 1–27). Lawrence Erlbaum Associates.

Bennett, R. E. (1998). Reinventing assessment: Speculations on the future of large-scale educational testing (Policy Information Perspective). ETS. http://www.ets.org/Media/Research/pdf/PICREINVENT

Bennett, R. E. (2011). Formative assessment: A critical review. Assessment in Education: Principles, Policy & Practice, 18(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. Educational Assessment, 28(2), 83–104. https://doi.org/10.1080/10627197.2023.2202312

Berman, A. I., Feuer, M. J., & Pellegrino, J. W. (2019). What use is educational assessment? The Annals of the American Academy of Political and Social Science, 683(1), 8–20. https://doi.org/10.1177/0002716219843871

Bernacki, M. L. (2018). Examining the cyclical, loosely sequenced, and contingent features of self-regulated learning: Trace data and their analysis. In D. H. Schunk & J. A. Greene (Eds.), Handbook of self-regulation of learning and performance (2nd ed., pp. 370–387). Routledge. https://doi.org/10.4324/9781315697048-24

Biddle, D. A., & Nooren, P. M. (2006). Validity generalization vs. Title VII: Can employers successfully defend tests without conducting local validation studies? Labor Law Journal, 57, 216–237. https://testgenius.com/articles/validity-generalization.pdf

Bicknell, K., Brust, C., & Settles, B. (2023, February 5). How Duolingo's AI learns what you need to learn. IEEE Spectrum. https://spectrum.ieee.org/duolingo

Bjork, E. L.,&Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough,&J. R. Pomerantz (Eds.), Psychology and the real world: Essays illustrating fundamental contributions to society (pp. 56–64). Worth Publishers.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7–74. https://doi.org/10.1080/0969595980050102

Blackman, R., & Ammanath, B. (2022, March 21). Ethics and AI: 3 conversations companies need to have. Harvard Business Review. https://hbr.org/2022/03/ethics-and-ai-3-conversations-companies-need-to-be-having

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13(6), 4–16. https://doi.org/10.3102/0013189X013006004

Bolsinova, M., Deonovic, B., Arieli-Attali, M., Burr, S., Hagiwara, M., & Maris, G. (2022). Measurement of ability in adaptive learning and assessment systems when learners use on-demand hints. Applied Psychological Measurement, 46(3), 219–235. https://doi.org/10.1177/01466216221084208

Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. Proceedings of the National Academy of Sciences, 115(26), 6674–6678. https://doi.org/10.1073/pnas.1718793115

Bresnahan, T. (2010). General purpose technologies. In B. H. Hall & N. Rosenberg (Eds.), Handbook of the economics of innovation (Vol. 2, pp. 761–791). https://doi.org/10.1016/S0169-7218(10)02002-2

Brookhart, S., Stiggins, R., McTighe, J., & Wiliam, D. (2020). The future of assessment practices: Comprehensive and balanced assessment systems. Learning Sciences International. https://testing123.education.mn.gov/cs/groups/communications/documents/document/mdaw/mdaw/~edis

Bradley, M. (1975). Scientific education versus military training: The influence of Napoleon Bonaparte on the Ecole Polytechnique. Annals of Science, 32(5), 415–449. https://doi.org/10.1080/00033797500200381

Buckley, J., Colosimo, L., Kantar, R., McCall, M., & Snow, E. (2021). Game-based assessment for education. In OECD digital education outlook 2021: Pushing the frontiers with artificial intelligence, blockchain and robots (pp. 195–208). OECD. https://read.oecd-ilibrary.org/education/oecd-digital-education-outlook-2021_9289cbfd-en#page1

Bull, S., & Kay, J. (2016). SMILI : A framework for interfaces to learning data in open learner models, learning analytics and related fields. International Journal of Artificial Intelligence in Education, 26, 293–331. https://doi.org/10.1007/s40593-015-0090-8

Burning Glass Technologies. (2019). Mapping the genome of jobs: The Burning Glass skills taxonomy [White paper]. https://www.voced.edu.au/content/ngv%3A84406

Burrus, J., Rikoon, S. H.,&Brenneman, M.W. (Eds.). (2022). Assessing competencies for social and emotional learning: Conceptualization, development, and applications. Routledge. https://doi.org/10.4324/9781003102243

BusinessWire. (2024). Carnegie learning wins 2024 EdTech award for MATHstream [Press release]. https://www.businesswire.com/news/home/20240327088407/en/Carnegie-Learning-Wins-2024-EdTech-Award-for-MATHstream

Buyse, T., & Lievens, F. (2011). Situational judgment tests as a new tool for dental student selection. Journal of Dental Education, 75(6), 743–749. https://doi.org/10.1002/j.0022-0337.2011.75.6.tb05101.x

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81–105. https://doi.org/10.1037/h0046016

cApStAn & Halleux, B. (2019). PISA 2021 translation and adaptation guidelines. OECD. https://www.oecd.org/pisa/pisaproducts/PISA-2022-Translation-and-Adaptation-Guidelines.pdf

Cao, M., Drasgow, F., & Cho, S. (2015). Developing ideal intermediate personality items for the ideal point model. Organizational Research Methods, 18(2), 252–275. https://doi.org/10.1177/1094428114555993

Casner-Lotto, J., & Barrington, L. (2006). Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century US workforce. Partnership for 21st Century Skills.

Cattell, R. B. (1965). A biometrics invited paper. Factor analysis: An introduction to essentials I. The purpose and underlying models. Biometrics, 21(1), 190–215. https://doi.org/10.2307/2528364

Cattell, R. B.,&Warburton, F.W. (1967). Objective personality and motivation tests: A theoretical introduction and practical compendium. University of Illinois Press.

Chakraborty, M., Tonmoy, T. I., Zaman, M., Gautam, S., Kumar, T., Sharma, K., Barman, N., Gupta, C., Jain, V., Chadha, A., Sheth, A.,& Das, A. (2023). Counter Turing test (CT2): AI-generated text detection is not as easy as you may think—Introducing AI detectability index (ADI). In H. Bouamor, J. Pino, & K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 2206–2239). ACL. https://aclanthology.org/2023.emnlp-main.136/

Cengage. (2019, January 16). New survey: demand for "uniquely human skills" increases even as technology and automation replace some jobs [Press release]. https://www.cengagegroup.com/news/press-releases/2019/new-survey-demand-for-uniquely-human-skillsincreases-even-as-technology-and-automation-replace-some-jobs/

Chamorro-Premuzic, T. (2021, May 26). The problem with job interviews. Forbes. https://www.forbes.com/sites/tomaspremuzic/2021/05/26/the-problem-with-job-interviews/?sh=4292b1224

Chan, S., Somasundaran, S., Ghosh, D., & Zhao, M. (2022). AGReE: A system for generating automated grammar reading exercises. In W. Che & E. Shutova (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 169–177). ACL. https://aclanthology.org/2022.emnlp-demos.17/

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. Journal of Economic Behavior & Organization, 149, 74–87. https://doi.org/10.1016/j.jebo.2018.02.024

Chen, L., Feng, G., Joe, J., Leong, C.W., Kitchen, C., & Lee, C. M. (2014). Towards automated assessment of public speaking skills using multimodal cues. In ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction (pp. 200–203). ACM. https://doi.org/10.1145/2663204.2663265

Chen, Y., Lee, Y.-H., & Li, X. (2022). Item pool quality control in educational testing: Change point model, compound risk, and sequential detection. Journal of Educational and Behavioral Statistics, 47(3), 322–352. https://doi.org/10.3102/10769986211059085

Cheng, K. H. C., Hui, C. H.,&Cascio, W. F. (2017). Leniency bias in performance ratings: The Big-Five correlates. Frontiers in Psychology, 8, Article 521. https://doi.org/10.3389/fpsyg.2017.00521

Chernyshenko, O. S., Kankaraš, M., & Drasgow, F. (2018). Social and emotional skills for student success and well-being: Conceptual framework for the OECD study on social and emotional skills (OECD Education Working Paper No. 173). OECD. https://one.oecd.org/document/EDU/WKP(2018)9/En/pdf

Chetty, R., Deming, D. J., & Friedman, J. N. (2023). Diversifying society's leaders? The determinants of causal effects of admission to highly selective private colleges (Working Paper No. 31492). National Bureau of Economic Research. https://doi.org/10.3386/w31492

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist, 49(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Choi, I., Hao, J., Deane, P.,&Zhang, M. (2021). Benchmark keystroke biometrics accuracy from high-stakes writing tasks (Research Report No. RR-21-15). ETS. https://doi.org/10.1002/ets2.12326

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. Personnel Psychology, 63(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Chopade, P., Edwards, D., Khan, S.M., Andrade, A.,&Pu, S. (2019, November). CPSX: using AI-machine learning for mapping human-human interaction and measurement of CPS teamwork skills. In 2019 IEEE International Symposium on Technologies for Homeland Security (HST) (pp. 1-6). IEEE.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213–220. https://doi.org/10.1037/h0026256

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

College Board. (2023, September 27). SAT suite: Everything you need to know about the Digital SAT. College Board Blog. https://blog.collegeboard.org/everything-you-need-know-about-digital-sat

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. Psychological Bulletin, 136(6), 1092–1122. https://doi.org/10.1037/a0021212

Cooper, W. H. (1981). Ubiquitous halo. Psychological Bulletin, 90(2), 218–244. https://doi.org/10.1037/0033-2909.90.2.218

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4, 253–278. https://doi.org/10.1007/BF01099821

Cotra, A. (2023, August 29). Language models surprised us. Planned Obsolescence. https://www.planned-obsolescence.org/language-models-surprised-us/

Cox, C. B., Barron, L. G., Davis, W., & de la Garza, B. (2017). Using situational judgment tests (SJTs) in training: Development and evaluation of a structured, low-fidelity scenario-based training method. Personnel Review, 46(1), 36–45. https://doi.org/10.1108/PR-05-2015-0137

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. American Psychologist, 30(1), 1–14. https://doi.org/10.1037/0003-066X.30.1.1

Darling-Hammond, L. (2001). Inequality in teaching and schooling: How opportunity is rationed to students of color in America. In B. D. Smedley, A. Y. Stith, L. Colburn, & C. H. Evans (Eds.), The right thing to do, the smart thing to do: Enhancing diversity in health professions—Summary of the Symposium on Diversity in Health Professions in Honor of Herbert W. Nickens, M. D. (pp. 208–233). National Academies Press. http://www.nap.edu/catalog/10186.html

Davey, T. (2023). Automated test assembly. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), International encyclopedia of education: Vol. 14. Quantitative research and educational measurement (pp. 201–208). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10027-2

Davoli, M., & Entorf, H. (2018). The PISA shock, socioeconomic inequality, and school reforms in Germany (IZA Policy Paper No. 140). IZA – Institute of Labor Economics. https://docs.iza.org/pp140.pdf

De Boeck, P. (2023, July 25–28). Pervasive DIF and DIF detection bias [Paper presentation]. International Meeting of the Psychometric Society (IMPS 2023), University of Maryland, College Park, MD, United States.

De Boeck, P., & Cho, S.-J. (2021). Not all DIF is shaped similarly. Psychometrika, 86(3), 712–716. https://doi.org/10.1007/s11336-021-09772-3

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. Review of Educational Research, 87(2), 243–282. https://doi.org/10.3102/0034654316687036

Dell. (2018, January 30). 3,800 business leaders declare: It's A tale of two futures. https://www.dell.com/en-us/perspectives/3800-business-leaders-declare-its-a-tale-of-two-futures/

Deming, D. J. (2017). The growing importance of social skills in the labor market. The Quarterly Journal of Economics, 132(4), 1593–1640. https://doi.org/10.1093/qje/qjx022

Deming, D. (2024, March 7). The worst way to do college admissions: Making standardized test scores optional has harmed the disadvantaged applicants it was intended to help. The Atlantic. https://theatlantic.com/ideas/archive/2024/03/standardized-testing-requirements-act-sat/677667/

Deming, D., & Kahn, L. B. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. Journal of Labor Economics, 36(S1), S337–S369. https://doi.org/10.1086/694106

Deonovic, B., Yudelson, M., Bolsinova, M., Attali, M.,&Maris, G. (2018). Learning meets assessment. Behaviormetrika, 45(2), 457–474. https://doi.org/10.1007/s41237-018-0070-z

Diao, Q., & van der Linden, W. J. (2013). Integrating test-form formatting into automated test assembly. Applied Psychological Measurement, 37(5), 361–374. https://doi.org/10.1177/0146621613476157

Di Battista, A., Grayling, S., Hasselaar, E., Leopold, T., Li, R., Rayner, M., & Zahidi, S. (2023, May). Future of jobs report 2023. World Economic Forum. https://www.weforum.org/reports/the-future-of-jobs-report-2023

DiCerbo, K. (2024, March 7). How we built AI tutoring tools. Khan Academy Blog. https://blog.khanacademy.org/how-we-built-ai-tutoring-tools/

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. Review of Educational Research, 87(2), 243–282. https://doi.org/10.3102/0034654316687036

Dobrescu, L., Holden, R., Motta, A., Piccoli A., Roberts, P., & Walker, S. (2021). Cultural context in standardized tests (Working Paper 2021-08). University of New South Wales Business School. https://doi.org/10.2139/ssrn.3983663

Duolingo Team. (2023, March 14). Introducing Duolingo Max, a learning experience powered by GPT-4. Duolingo Blog. https://blog.duolingo.com/duolingo-max/

Eberly Center. (n.d.). Learning principles: Theory and research-based principles of learning. Carnegie Mellon University. https://www.cmu.edu/teaching/principles/learning.html

Elliott, S.W. (2017). Computers and the future of skill demand. OECD. https://doi.org/10.1787/978926428439 en

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv. https://arxiv.org/abs/2303.10130v4

Embretson, S. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed), Cognitive assessment: A multidisciplinary perspective (pp. 107–135). Springer.

Emerson, A., Houghton, P., Chen, K., Basheerabad, V., Ubale, R., Leong, C.W. (2022). Predicting user confidence in video recordings with spatio-temporal multimodal analytics. In ICMI '22 companion: Companion publication of the 2022 International Conference on Multimodal Interaction (pp. 98–104). ACM. https://doi.org/10.1145/3536220.3558007

Erwin, T. D., & Sebrell, K. W. (2003). Assessment of critical thinking: ETS's tasks in critical thinking. Journal of General Education, 52(1), 50–70. https://doi.org/10.1353/jge.2003.0019

ETS. (n.d.). Demonstrate program effectiveness with the ETS® Major Field Tests. https://www.ets.org/mft.html

ETS. (2014). ETS standards for quality and fairness. https://ets.org/pdfs/about/standards-quality-fairness.pdf

ETS. (2022). ETS guidelines for developing fair tests and communications. https://www.ets.org/pdfs/about/fai tests-andcommunications.pdf

ETS. (2023a). ETS human progress study [Unpublished data set].

ETS. (2023b). Your at home testing. https://www.ets.org/gre/test-takers/general-test/register/at-home-testing.html

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. The Quarterly Journal of Economics, 133(4), 1645–1692. https://doi.org/10.1093/qje/qjy013

Feuer, M. J. (2012). No country left behind: Rhetoric and reality of international large-scale assessment. ETS. http://www.ets.org/Media/Research/pdf/PICANG13.pdf

Feuer, M., Holland, P.W., Green, B. F., Bertenthal, M.W.,&Hemphill, F. C. (Eds.). (1999). Uncommon measures: Equivalence and linkage among educational tests. National Academies Press. https://doi.org/10.17226/6332

Flanagan, C. (2021, July 22). The University of California is lying to us. The Atlantic. https://www.theatlantic.com/ideas/archive/2021/07/why-university-california-dropping-sat/619522/

Flynn, M. (2023, May 30). The soft skills "debate" is over. Forbes. https://www.forbes.com/sites/mariaflynn/20 soft-skills-debate-is-over/?sh=5baa274b7308

Foster, N., & Piacentini, M. (Eds.). (2023). Innovating assessments to measure and support complex skills. OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Frensch, P. A., & Funke, J. (1995). Complex problem solving: The European perspective. Routledge.

Frey, C. B.,&Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerization? Technological Forecasting and Social Change, 114, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Friedland, N. S., Allen, P. G., Matthews, G., Witbrock, M., Baxter, D., Curtis, J., Shepard, B., Miraglia, P., Angele, J., Staab, S., Moench, E., Oppermann, H., Wenke, D., Israel, D., Chaudhri, V., Porter, B., Barker, K., Fan, J., Chaw, S., … Clark, P. (2004). Project Halo: Towards a digital Aristotle. AI Magazine, 25(4), 29–47. https://doi.org/10.1609/aimag.v25i4.1783

Fu, J., Kyllonen, P. C., & Tan, X. (2024). From Likert to forced choice: Statement parameter invariance and context effects in personality assessment. Measurement: Interdisciplinary Research and Perspectives. Advance online publication. https://doi.org/10.1080/15366367.2023.2258482

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. Exceptional Children, 53(3), 199–208. https://doi.org/10.1177/001440298605300301

Fyfe, E. R., Borriello, G. A., & Merrick, M. (2023). A developmental perspective on feedback: How corrective feedback influences children's literacy, mathematics, and problem solving. Educational Psychologist, 58(3), 130–145. https://doi.org/10.1080/00461520.2022.2108426

Fyfe, E. R., De Leeuw, J. R., Carvalho, P. F., Goldstone, R. L., Sherman, J., Admiraal, D., Alford, L.K., Bonner, A., Brassil, C. E., Brooks, C. A., Carbonetto, T., Chang, S. H., Cruz, L., Czymoniewicz-Klippel, Daniel, F., Driessen, M., Habashy, N., Hanson-Bradley, C. L., Hirt, E. R., … Motz, B. A. (2021). Many Classes 1: Assessing the generalizable effect of immediate feedback versus delayed feedback across many college classes. Advances in Methods and Practices in Psychological Science, 4(3), Article 25152459211027575. https://doi.org/10.1177/25152459211027575

Gao, L., Ghosh, D.,&Gimpel, K. (2022). What makes a question inquisitive? A study on type-controlled inquisitive question generation. In V. Nastase, E. Pavlick, M.T. Pilehvar, J. Camacho-Callados,&A. Raganato (Eds.), Proceedings of the 11th Joint Conference on Lexical and Computational Semantics (pp. 240–257). ACL. https://doi.org/10.18653/v1/2022.starsem-1.22

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2414–2423). IEEE. https://doi.org/10.1109/CVPR.2016.265

Geerlings, H., Glas, C.A.,&Van Der Linden, W. J. (2011). Modeling rule-based item generation. Psychometrika, 76, 337–359. https://doi.org/10.1007/s11336-011-9204-x

Geiger, M., Bärwaldt, R., & Wilhelm, O. (2021). The good, the bad, and the clever: Faking ability as a socio-emotional ability? Journal of Intelligence, 9(1), 1–22. https://doi.org/10.3390/jintelligence9010013

Geisinger, K. F. (2011). The future of high-stakes testing in education. In J. A. Bovaird, K. F. Geisinger, & C.W. Buckendahl (Eds.), High-stakes testing in education: Science and practice in K–12 settings (pp. 231–248). American Psychological Association. https://doi.org/10.1037/12330-014

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). Automatic item generation: Theory and practice. Routledge.

Gil, Y., & Selman, B. (2019). A 20-year community roadmap for artificial intelligence research in the US. arXiv. https://doi.org/10.48550/arXiv.1908.02624

Glas, C. A. W., & van der Linden, W. J. (2001, June 2–4). Modeling variability in item parameters in CAT [Paper presentation]. North American Psychometric Society Meeting King of Prussia, PA, United States.

Godwin, K. E., Almeda, M. V., Seltman, H., Kai, S., Skerbetz, M. D., Baker, R. S., & Fisher, A. V. (2016). Off-task behavior in elementary school children. Learning and Instruction, 44, 128–143. https://doi.org/10.1016/j.learninstruc.2016.04.003

Goldberg, B., & Sinatra, A. M. (2023). Generalized intelligent framework for tutoring (gift) SWOT analysis. In A. M. Sinatra, A. C. Graesser, X. Hu, G. Goodwin, & V. Rus (Eds.), Design recommendations for intelligent tutoring systems: Vol. 10. Strengths, weaknesses, opportunities and threats (SWOT) analysis of intelligent tutoring systems (pp. 9–26). U.S. Army Combat Capabilities Development Command—Soldier Center. https://gifttutoring.org/documents/163

Goodhart, C. A. E. (1984). Monetary theory and practice: The U.K. experience. Springer. https://doi.org/10.1007/978-1-349-17295-5

The Gordon Commission on the Future of Assessment in Education. (2013). To assess, to teach, to learn: A vision for the future of assessment. ETS. https://www.ets.org/Media/Research/pdf/gordon_c

Gosling, S. D., Augustine, A. A., Vazire, S., Holtzman, N., & Gaddis, S. (2011). Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information. Cyberpsychology, Behavior, and Social Networking, 14(9), 483–488. https://doi.org/10.1089/cyber.2010.0087

Gosling, S.D., Ko, S. J., Mannarelli, T.,&Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. Journal of Personality and Social Psychology, 82(3), 379–398. https://doi.org/10.1037//0022-3514.82.3.379

Graf, E. A., & Fife, J. H. (2012). Difficulty modeling and automatic generation of quantitative items: Recent advances and possible next steps. In M. J. Gierl & T. M. Haladyna (Eds.), Automatic item generation (pp. 157–178). Routledge.

Greiff, S., Gaševi´c, D., & von Davier, A. (2017). Using process data for assessment in intelligent tutoring systems: A cognitive psychologist, psychometrician, and computer scientist perspective. In R. Sottilare, A. Graesser, X. Hu, & G. Goodwin (Eds.), Design recommendations for intelligent tutoring systems: Vol. 5. Assessment methods (pp. 171–179). U.S. Army Research Laboratory. https://gifttutoring.org/attachments/download/2410/Design%20Recommendations%2
%20Assessment_final_errata%20corrected.pdf

Grigorenko, E. L.,&Sternberg, R. J. (1998). Dynamic testing. Psychological Bulletin, 124(1), 75–111. https://doi.org/10.1037/0033-2909.124.1.75

Grose, J. (2024, January 17). Don't ditch standardized tests: Fix them. The New York Times. https://www.nytimes.com/2024/01/17/opinion/standardized-tests.html

Grossmann, I., Rotella, A. Sharpinskyi, K., Browne, D. T., & Fong, G. T. (2023). Insights into the accuracy of social scientists' forecasts of societal change. Nature Human Behavior, 7, 484–501. https://doi.org/10.1038/s41562-022-01517-1

Haberman, S. J., & Lee, Y.-H. (2017). A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses (Research Report No. RR-17-23). ETS. https://doi.org/10.1002/ets2.12150

Haberman, S. J., Lee, Y.-H., Papierman, P., Zhou, Y., & Subhedar, R. (2022). Systems and methods for detecting unusually frequent exactly matching and nearly matching test responses (U.S. Patent 11,398,161). U.S. Patent Office and Trademark Office. https://ppubs.uspto.gov/pubwebapp/external.html?q=(11398161).pn.&db=USPAT&type=ids

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In National Research Council (Ed), Methodological advances in cross-national surveys of educational achievement (pp. 58–79). National Academies Press. https://nap.nationalacademies.org/read/10322/chapter/4

Hao, J., Liu, L., Kyllonen, P. C., Flor, M., & von Davier, A. A. (2019). Psychometric considerations and a general scoring strategy for assessments of collaborative problem solving (Research Report No. RR-19-41). ETS. https://doi.org/10.1002/ets2.12276

Hao, J., Liu, L., von Davier, A. A., Lederer, N., Zapata-Rivera, D., Jakl, P.,&Bakkenson, M. (2017). EPCAL: ETS platform for collaborative assessment and learning (Research Report No. RR-17-49). ETS. https://doi.org/10.1002/ets2.12181

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: the impacts and implications of large language models and generative AI. Educational Measurement: Issues and Practices. Advance online publication.

Hattie, J. A. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge.

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In E. Mayer & P. A. Alexander (Eds.), Handbook of research on learning and instruction (pp. 249–271). Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

He, J., Bartram, D., Inceoglu, I., & van de Vijver, F. J. R. (2014). Response styles and personality traits: A multilevel analysis. Journal of Cross-Cultural Psychology, 45(7), 1028–1045. https://doi.org/10.1177/0022022114534773

He, Q., Borgonovi, F., & Paccagnella, M. (2019). Using process data to understand adults' problem-solving behaviour in the programme for the international assessment of adult competencies (PIAAC): Identifying generalised patterns across multiple tasks with sequence mining (OECD Education working paper No. 205). OECD. https://one.oecd.org/document/EDU/WKP(2019)13/en/pdf

Heckman, J.,&Zhou, J. (2021). Interactions as investments: The microdynamics and measurement of early childhood learning [Manuscript submitted for publication].

Hedlund, J., Wilt, J. M., Nebel, K. L., Ashford, S. J., & Sternberg, R. J. (2006). Assessing practical intelligence in business school admissions: A supplement to the graduate management admissions test. Learning and Individual Differences, 16(2), 101–127. https://doi.org/10.1016/j.lindif.2005.07.005

Herman, J. L., Martínez, J. F.,&Bailey, A. L. (2023). Fairness in educational assessment and the next edition of the standards: Concluding commentary. Educational Assessment, 28(2), 128–136. https://doi.org/10.1080/10627197.2023.2215980

Hilton, M., & Herman, J. (Eds.). (2017). Supporting students' college success: The role of assessment of intrapersonal and interpersonal competencies. National Academies Press.

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. Journal of Chiropractic Education, 33(2), 151–163. https://doi.org/10.7899/JCE-18-22

Hinnant-Crawford, B. N. (2020). Improvement science in education: A primer. Myers Education Press.

Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. Economics of Education Review, 52, 105–119. https://doi.org/10.1016/j.econedurev.2016.02.001

Holland, P.W. (1996). Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support (Research Report No. RR-96-07). ETS. https://doi.org/10.1002/j.2333-8504.1996.tb01685.x

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. Journal of Negro Education, 67(3), 187–196. https://doi.org/10.2307/2668188

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. Psychological Methods, 4(4), 403–424. https://doi.org/10.1037/1082-989X.4.4.403

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017). Toward controlled generation of text. In D. Precup & Y. W. Teh (Eds.), Proceedings of machine learning research: Vol. 70. Proceedings of the 34th International Conference on Machine Learning (pp. 1587–1598). https://proceedings.mlr.press/v70/hu17e.html

IMS Global. (2022). Question & test interoperability (QTI) 3.0: Best practices and implementation guide. https://www.imsglobal.org/spec/qti/v3p0/impl/

Institute of Medicine. (2015). Psychological testing in the service of disability determination. The National Academies Press. https://doi.org/10.17226/21704

International Test Commission. (2001). International guidelines for test use. International Journal of Testing, 1(2), 93–114. https://doi.org/10.1207/S15327574IJT0102_1

International Test Commission. (2013). ITC guidelines for test use. Final version. https://www.intestcom.org/files/guideline_test_use.pdf

International Test Commission. (2017). The ITC guidelines for translating and adapting tests (2nd ed.). https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

International Test Commission & Association of Test Publishers. (2022). Guidelines for technology-based assessment. https://www.intestcom.org/upload/media-library/guidelines-for-technology-based-assessment-v20221108-16684036687NAG8.pdf

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2013). Item generation for test development. Routledge.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non-test score outcomes. Journal of Political Economy, 126(5), 2072–2107. https://doi.org/10.1086/699018

Jiang, Y., Martin-Raugh, M., Yang, Z., Hao, J., Liu, L., & Kyllonen, P. C. (2023). Do you know your partner's personality through virtual collaboration or negotiation? Investigating perceptions of personality and their impacts on performance. Computers in Human Behavior, 141, Article 107608. https://doi.org/10.1016/j.chb.2022.107608

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), Handbook of personality: Theory and research (Vol. 2., pp. 102–138). Guilford Press

Johnson, M. S. (2024). How do we demonstrate AI responsibility: The devil is in the details. [Manuscript in preparation].

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. Journal of Educational Measurement, 59(3), 338–361. https://doi.org/10.1111/jedm.12335

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. In S. Lane (Ed.), Advancing natural language processing in educational assessment (pp. 143–164). Routledge. https://doi.org/10.4324/9781003278658-12

Johnson, M. S., & Sinharay, S. (2005). Calibration of polytomous item families using Bayesian hierarchical modeling. Applied Psychological Measurement, 29(5), 369–400. https://doi.org/10.1177/0146621605276675

Jung, J. Y., Tyack, L., & von Davier, M. (2022). Automated scoring of constructed-response items using artificial neural networks in international large-scale assessment. Psychological Test and Assessment Modeling, 64(4), 471–494.

Karay, Y., Reiss, B.,&Schauber, S. K. (2020). Progress testing anytime and anywhere: Does a mobile-learning approach enhance the utility of a large-scale formative assessment tool? Medical Teacher, 42(10), 1154–1162. https://doi.org/10.1080/0142159X.2020.1798910

Karpicke, J.D.,&Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. Science, 331(6018), 772–775. https://doi.org/10.1126/science.1199327

Kautz, T., & Zanoni, W. (2014). Measuring and fostering non-cognitive skills in adolescence: Evidence from Chicago public schools and the OneGoal program. University of Chicago.

Kell, H. J., Martin-Raugh, M. P., Carney, L. M., Inglese, P. A., Chen, L., & Feng, G. (2017). Exploring methods for developing behaviorally anchored rating scales for evaluating structured interview performance (Research Report No. RR-17-28). ETS. https://doi.org/10.1002/ets2.12152

Kessler, J. B., Low, C., & Sullivan, C. D. (2019). Incentivized resume rating: Eliciting employer preferences without deception. American Economic Review, 109(11), 3713–3744. https://doi.org/10.1257/aer.20181714

King, G.,&Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. Political Analysis, 15(1), 46–66. https://doi.org/10.1093/pan/mpl011

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. Educational Measurement: Issues and Practice, 30(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Klieger, D. M., Kell, H. J., Rikoon, S., Burkander, K. N., Bochenek, J. L., & Shore, J. R. (2018). Development of the behaviorally anchored rating scales for the skills demonstration and progression guide (Research Report No. RR-18-24). ETS. https://doi.org/10.1002/ets2.12210

Klinger, D. A., McDivitt, P. R., Howard, B. B., Munoz, M. A., Rogers, W. T., & Wylie, E. C. (2015). The classroom assessment standards for preK-12 teachers. Kindle Direct Press.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin, 119(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). Formative assessment and elementary school student academic achievement: A review of the evidence (Report No. REL 2017-259). Regional Educational Laboratory Central.

Koedinger, K. R., Carvalho, P. F., Liu, R., & McLaughlin, E. A. (2023). An astonishing regularity in student learning rate. Proceedings of the National Academy of Sciences, 120(13), Article e2221311120. https://doi.org/10.1073/pnas.2221311120

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks. Machine Learning, 95, 357–380. https://doi.org/10.1007/s10994-013-5415-y

Krachman, S. B., Arnold, R., & LaRocca, R. (2016). Expanding the definition of student success: A case study of the CORE districts. Transforming Education. https://transformingeducation.org/wp-content/uploads/2017/04/TransformingEducationCaseStudyFINAL1.p

Kukea Shultz, P., & Englert, K. (2021). Cultural validity as foundational to assessment development: An indigenous example. Frontiers in Education, 6, Article 701973. https://doi.org/10.3389/feduc.2021.701973

Kulik, J. A.,&Fletcher, J.D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. Review of Educational Research, 86(1), 42–78. https://doi.org/10.3102/0034654315581420

Kumar, V., & Boulanger, D. (2020, October). Explainable automated essay scoring: Deep learning really has pedagogical value. In Frontiers in Education, 5, Article 572367. https://doi.org/10.3389/feduc.2020.572367

Kuncel, N. R., Kochevar, R. J., & Ones, D. S. (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. International Journal of Selection and Assessment, 22(1), 101–107. https://doi.org/10.1111/ijsa.12060

Kyllonen, P. C. (2016). Socio-emotional and self-management variables in learning and assessment. In A. A. Rupp & J. P. Leighton (Eds.), The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications (pp. 174–197). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch8

Kyllonen, P. (2021). Taxonomy of cognitive abilities and measures for assessing artificial intelligence and robotics capabilities. In AI and the future of skills: Volume 1. Capabilities and assessments (pp. 50–76). OECD Publishing. https://doi.org/10.1787/feecd512-en

Kyllonen, P. C., & Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski

(Eds.), Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 277–285). CRC Press.

Kyllonen, P., Hao, J., Weeks, J., Fauss, M., & Kerzabi, E. (2023). Collaborative problem solving (CPS) skill: Estimating an individual's contribution to small group performance [Unpublished manuscript]. ETS.

Kyllonen, P., Hartman, R., Sprenger, A., Weeks, J., Bertling, M., McGrew, K., Kriz, S., Bertling, J., Fife, J., & Stankov, L. (2019). General fluid/inductive reasoning battery for a high-ability population. Behavior Research Methods, 51(2), 507–522. https://doi.org/10.3758/s13428-018-1098-4

Kyllonen, P. C., & Kell, H. (2018). Ability tests measure personality, personality tests measure ability: Disentangling construct and method in evaluating the relationship between personality and ability. Journal of Intelligence, 6(3), Article 32, https://doi.org/10.3390/jintelligence6030032

Kyriazos, T. A. (2018). Applied psychometrics: The application of CFA to multitrait-multimethod matrices (CFA-MTMM). Psychology, 9(12), 2625–2648. https://doi.org/10.4236/psych.2018.9121

Landers, R. N., Armstrong, M. B., Collmus, A. B., Mujcic, S., & Blaik, J. (2022). Theory-driven game-based assessment of general cognitive ability: Design theory, measurement, prediction of performance, and test fairness. Journal of Applied Psychology, 107(10), 1655–1677. https://doi.org/10.1037/apl0000954

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

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). Handbook of test development (Vol. 2, pp. 3–18). Routledge.

Lang, J. W. B., & Tay, L. (2021). The science and practice of item response theory in organizations. Annual Review of Organizational Psychology and Organizational Behavior, 8, 311–338. https://doi.org/10.1146/annurev-orgpsych-012420-061705

Langer, C., & Wiederhold, S. (2023). The value of early-career skills (CESifo Working Paper No. 10288). CESifo Network. https://doi.org/10.2139/ssrn.4369987

Lassébie, J., & Quintini, G. (2022). What skills and abilities can automation technologies replicate and what does it mean for workers? New evidence (OECD Social, Employment and Migration Working Papers, No. 282). OECD Publishing. https://doi.org/10.1787/646aad77-en

Law, K. S., Mobley, W. H., & Wong, C.-S. (2002). Impression management and faking in biodata scores among Chinese job-seekers. Asia Pacific Journal of Management, 19, 541–556. https://doi.org/10.1023/A:1020521726390

Lederman, O., Calacci, D., MacMullen, A., Fehder, D. C., Murray, F. E., & Pentland, A.S. (2016). Open badges: A low-cost toolkit for measuring team communication and dynamics. In The online proceedings of the 2016 International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BriMS 2016). http://sbp-brims.org/2016/proceedings/IN_105.pdf

Lee, G. H., Lee, K. J., Jeong, B. & Kim, T. (2024). Developing personalized marketing service using generative AI. IEEE Access, 12, 22394–22402. https://doi.org/10.1109/ACCESS.2024.3361946

Lee, H. A. (2023, January 23). This is why Microsoft Kinect was a complete failure. SVG. https://www.svg.com/301470/this-is-why-microsoft-kinect-was-a-complete-failure/

Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. Psychometrika, 78(4), 815–829. https://doi.org/10.1007/s11336-013-9337-1

Lee, Y.-H., & Haberman, S. J. (2021). Studying score stability with a harmonic regression family: A comparison of three approaches to adjustment of examinee-specific demographic data. Journal of Educational Measurement, 58(1), 54–82. https://doi.org/10.1111/jedm.12266

Lee, Y.-H., & Lewis, C. (2021). Monitoring item performance with CUSUM statistics in continuous testing. Journal of Educational and Behavioral Statistics, 46(5), 611–648. https://doi.org/10.3102/1076998621994563

Lee, Y.-H., Lewis, C., & von Davier, A. A. (2014). Monitoring the quality and security of multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), Computerized multistage testing: Theory and applications (pp. 285–300). CRC Press.

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. Psychometrika, 78(3), 557–575. https://doi.org/10.1007/s11336-013-9317-5

Leenknecht, M., Hompus, P., & van der Schaaf, M. (2019). Feedback seeking behaviour in higher education: The association with students' goal orientation and deep learning approach. Assessment & Evaluation in Higher Education, 44(7), 1069–1078. https://doi.org/10.1080/02602938.2019.1571161

Lehman, B., Sparks, J. R., & Zapata-Rivera, D. (2018). When should an adaptive assessment care? In N. Guin & A. Kumar (Eds.), Proceedings of ITS 2018: Intelligent Tutoring Systems 14th International Conference, Workshop on Exploring Opportunities for Caring Assessments (pp. 87–94). ITS. https://ceur-ws.org/Vol-2354/w3paper1.pdf

Leonhardt, D. (2024, January 7). The misguided war on the SAT. The New York Times. https://www.nytimes.com/2024/01/07/briefing/the-misguided-war-on-the-sat.html

Lewin, T. (2002, December 4). Henry Chauncey dies at 97; Shaped admission testing for the nation's colleges. The New York Times. https://www.nytimes.com/2002/12/04/nyregion/henry-chauncey-dies-at-97-shaped-admission-testing-for-the-nation-s-colleges.html

Lewis, C. (2001). Expected response functions. In A. Boomsma, M. A J. van Duijn, & T. A. B. Snijders (Eds.), Essays on item response theory (pp. 163–171). Springer. https://doi.org/10.1007/978-1-4613-0169-1_9

Lewis, C.,&Thayer, D. T. (1998). The power of the K-index (or PMIR) to detect copying (Research Report No. RR-98-49). ETS. https://doi.org/10.1002/j.2333-8504.1998.tb01798.x

LinkedIn Talent Solutions. (2019). Global talent trends: The 3 trends transforming your workplace. https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/global_talent_trends_2019_emea.pdf

Linzarini, A., & Catarino da Silva, D. (2024). Innovative assessments for Social Emotional Skills [webinar slides]. SlideShare. https://www.slideshare.net/slideshow/webinar-innovative-assessments-for-social-emotional-skills/270083576

Lira, B., O'Brien, J. M., Peña, P. A., Galla, B. M., D'Mello, S., Yeager, D. S., Defnet, A., Kautz, T., Munkacsy, K., & Duckworth, A. L. (2022). Large studies reveal how reference bias limits policy applications of self-report measures. Scientific Reports, 12, Article 19189. https://doi.org/10.1038/s41598-022-23373-9

Lissitz, R.W. (2009). Introduction. In R.W. Lissitz (Ed.), The concept of validity: Revisions, new directions, and applications (pp. 1–15). IAP Information Age Publishing.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. Educational Researcher, 41(9), 352–362. https://doi.org/10.3102/0013189X12459679

Liu, O. L., Kell, H. J., Liu, L., Ling, G., Wang, Y., Wylie, C., Sevak, A., Sherer, D., LeMahieu, P., & Knowles, T. (2023). A new vision for skills-based assessment. ETS. https://ets.org/pdfs/rd/new-vision-skills-based-assessment.pdf

Liu, O. L., Mao, L., Frankel, L., & Xu, J. (2016). Assessing critical thinking in higher education: The HEIghten approach and preliminary validity evidence. Assessment & Evaluation in Higher Education, 41(5), 677–694. https://doi.org/10.1080/02602938.2016.1168358

Liu, X., Zhang, Z., Wang, Y., Pu, H., Lan, Y., & Shen, C. (2023). COCO: Coherence-enhanced machine-generated text detection under low resource with contrastive learning. In H. Bouamor, J. Pino, & K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 16167–16188). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.1005

Loewus, L. (2016). What is digital literacy? Education Week. https://www.edweek.org/teaching-learning/what-is-digital-literacy/2016/11

Loukina, A., Yoon, S.-Y., Sakano, J., Wei, Y., & Sheehan, K. (2016). Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In Y. Matsumoto & R. Prasad (Eds.), Proceedings of COLING 2016, the 26th International Conference on

Computational Linguistics: Technical papers (pp. 3245–3253). https://aclanthology.org/C16-1306

Ludlow, L. H., O'Keefe, T., Braun, H., Anghel, E., Szendey, O., Matz, C., & Howell, B., (2022). An enhancement to the theory and measurement of purpose. Practical Assessment, Research, and Evaluation 27(1), Article 4. https://doi.org/10.7275/c5jb-rr95

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. Journal of Educational Psychology, 106(4), 901–918. https://doi.org/10.1037/a0037123

MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. Emotion, 8(4), 540–551. https://doi.org/10.1037/a0012746

Madnani, N., & Cahill, A. (2018). Automated scoring: Beyond natural language processing. In E. M. Bender, L.Derczynski, & P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics (pp. 1099–1109). ACL. https://aclanthology.org/C18-1094

Mammadov, S. (2022). Big Five personality traits and academic performance: A meta-analysis. Journal of Personality, 90(2), 222–255. https://doi.org/10.1111/jopy.12663

Mankki, V. (2023). Research using teacher or teacher educator job advertisements: A scoping review. Cogent Education, 10(1), Article 2223814. https://doi.org/10.1080/2331186X.2023.2223814

Martin-Raugh, M. P., Kyllonen, P. C., Hao, J., Bacall, A., Becker, D., Kurzum, C., Yang, Z., Yan, F., & Barnwell, P. (2020). Negotiation as an interpersonal skill: Generalizability of negotiation outcomes and tactics across contexts at the individual and collective levels. Computers in Human Behavior, 104, Article 105966. https://doi.org/10.1016/j.chb.2019.03.030

Martín-Raugh, M., Roohr, K. C., Leong, C. W., Molloy, H., McCulla, L., Ramanarayan, V., & Mladineo, Z. (2023). Better understanding oral communication skills: The impact of perceived personality traits. American Journal of Distance Education. Advance online publication. https://doi.org/10.1080/08923647.2023.2235950

Mattingly, S.M., Gregg, J.M., Audia, P., Bayraktaroglu, A. E., Campbell, A. T., Chawla, N. V., Das Swain, V., DeChoudhury, M., D'Mello, S. K., Dey, A. K., Gao, G., Jagannath, K., Jiang, K., Lin, S., Liu, Q., Mark, G., Martinez, G. J. Masaba, K., Mirjafari, S., … Striege, A. (2019, May). The tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers. In Extended abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-8). ACM. https://doi.org/10.1145/3290607.3299041

McLaughlin, K., Ainslie. M., Coderre, S., Wright, B., & Violato, C. (2009). The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. Medical Education, 43(10), 989–992. https://10.1111/j.1365-2923.2009.03438.x

McWhorter, J. (2024, March 14). No, the SAT isn't racist. The New York Times. https://www.nytimes.com/2024/03/14/opinion/sat-college-admissions-antiracism.html

Mervosh, S. (2022, September 1). The pandemic erased two decades of progress in math and reading: The results of a national test showed just how devastating the last two years have been for 9-year-old schoolchildren, especially the most vulnerable. The New York Times. https://www.nytimes.com/2022/09/01/us/national-test-scores-math-reading-pandemic.html

Meyer, R. H., Wang, C., & Rice, A. B. (2018). Measuring students' social-emotional learning among California's CORE districts: An IRT modeling approach [Working paper]. Policy Analysis for California Education. https://edpolicyinca.org/sites/default/files/Measuring_SEL_May-2018.pdf

Mignogna, G., Carey, C. E., Wedow, R., Baya, N., Cordioli, M., Pirastu, N., Bellocco, R., Mlerbi, K. F., Nivard, M. G., Neale, B. M., Walters, R. K., & Ganna, A. (2023). Patterns of item nonresponse behaviour to survey questionnaires are systematic and associated with genetic loci. Nature Human Behaviour, 7, 1371–1387. https://doi.org/10.1038/s41562-023-01632-7

Millsap, R. (2011). Statistical approaches to measurement invariance. Routledge.

Mirjafari, S., Masaba, K., Grover, T., Wang, W., Audia, P., Campbell, A. T., Chawla, N. V., Das Swain, V., De Choudhury, M., Dey, A. K., D'Mello, S. K., Gao, G., Gregg, J. M., Jagannath, K., Jiang, K., Lin, S., Qiang, L., Mark, G., Martinez, G. J., Martinez, S. M., … Striegel, A. (2019). Differentiating higher and lower job performers in the workplace using mobile sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3(2), 1–24. https://doi.org/10.1145/3328908

Mislevy, R. (2018). Sociocognitive foundations of educational measurement. Routledge.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K., & Michael, J. (2014). Psychometric considerations in game-based assessment. GlassLab Research, Institute of Play. https://web.archive.org/web/20160320151604/http://www.instituteofplay.org/wp-content/uploads/2014/02/GlassLab_GBA1_WhitePaperFull.pdf

Mislevy, R. J., Sheehan, K.M., & Wingersky, M. (1993). How to equate tests with little or no data. Journal of Educational Measurement, 30(1), 55–78. https://doi.org/10.1111/j.1745-3984.1993.tb00422.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Molenaar, I., de Mooij, S., Azevedo, R., Bannert, M., Järvelä, S., & Gaševi´c, D. (2023). Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data. Computers in Human Behavior, 139, Article 107540. https://doi.org/10.1016/j.chb.2022.107540

Morell, Z. (2017). Introduction to the New York State next generation early learning standards. https://www.nysed.gov/sites/default/files/introduction-to-the-nys-early-learning-standards.pdf

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. Instructional Science, 32(1–2), 99–113. https://doi.org/10.1023/B:TRUC.0000021811.66966.1d

Moro, E., Frank, M. R., Pentland, A., Rutherford, A., Cebrian, M., & Rahwan, I. (2021). Universal resilience patterns in labor markets. Nature Communications, 12, Article 1972. https://doi.org/10.1038/s41467-021-22086-3

Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. Applied Psychological Measurement, 11(1), 1–31. https://doi.org/10.1177/014662168701100101

Murphy, S. C., Klieger, D. M., Borneman, M. J., & Kuncel, N. R. (2009). The predictive power of personal statements in admissions: A meta-analysis and cautionary tale. College and University, 84(4), 83–86.

Narciss, S. (2004). The impact of informative tutoring feedback and self-efficacy on motivation and achievement in concept learning. Experimental Psychology, 51(3), 214–228. https://doi.org/10.1027/1618-3169.51.3.214

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. Computers & Education, 71, 56–76. https://doi.org/10.1016/j.compedu.2013.09.011

National Academies of Sciences, Engineering, and Medicine. (2018). How people learn II: Learners, contexts, and cultures. The National Academies Press. https://doi.org/10.17226/24783

National Academies of Sciences, Engineering, and Medicine. (2019). Monitoring educational equity. The National Academies Press. https://doi.org/10.17226/25389

National Association of Colleges and Employers. (2022). NACE job outlook 2022. https://www.naceweb.org/uploadedFiles/files/2022/resources/nace-job-outlook-2022.pdf

National Research Council. (1999a). High stakes: Testing for tracking, promotion, and graduation. The National Academies Press. https://doi.org/10.17226/6336

National Research Council. (1999b). Myths and tradeoffs: The role of tests in undergraduate admissions. The National Academies Press. https://doi.org/10.17226/9632

National Research Council. (2000). How people learn: Brain, mind, experience, and school (expanded ed.). The National Academies Press. https://doi.org/10.17226/9853

National Research Council (2001). Knowing what students know: The science and design of educational assessment. The National Academies Press. https://doi.org/10.17226/10019.

National Research Council. (2012). Education for life and work: Developing transferable knowledge and skills in the 21st century. The National Academies Press. https://doi.org/10.17226/13398.

Nesbit, J.C., Adesope, O.O., Liu, Q.,&Ma, W. (2014, July). How effective are intelligent tutoring systems in computer science education? In 2014 IEEE 14th International Conference on Advanced Learning Technologies (pp. 99–103). IEEE. https://doi.org/10.1109/ICALT.2014.38

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome. Measurement, 7(1), 23–35. https://doi.org/10.1007/s40271-013-0041-0

Nickow, A., Oreopoulos, P., & Quan, V. (2020). The impressive effects of tutoring on PreK-12 learning: A systematic review and meta-analysis of the experimental evidence (NBER working paper No. 27476). National Bureau of Economic Research. https://doi.org/10.3386/w27476

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017). Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. Personality and Individual Differences, 106, 183–189. https://doi.org/10.1016/j.paid.2016.11.014

Noor, N., Beram, S., Yuet, F. K. C., Gengatharan, K., Syafiq, M., & Rasidi, M. S. M. (2023). Bias, halo effect and horn effect: A systematic literature review. International Journal of Academic Research in Business & Social Sciences, 13(3), 1116–1140. https://doi.org/10.6007/IJARBSS/v13-i3/16733

Norville, V. (2022). States sketch 'portraits of a graduate.' State Innovations, 27(1), 1–4.

Novarese, M., & Di Giovinazzo, V. (2013). Promptness and academic performance (MPRA Paper No. 49746). Munich Personal RePEc Archive. https://mpra.ub.uni-muenchen.de/49746/

Ober, T. M., Lehman, B. A., Gooch, R., Oluwalana, O., Solyst, J., Phelps, G., & Hamilton, L. S. (2023). Culturally responsive learning: Recommendations for a working definition and framework (Research Report No. RR-23-09). Educational Testing Service. https://doi.org/10.1002/ets2.12372

O'Dwyer, E., Sparks, J. R., & Nabors Oláh, L. (2023). Enacting a process for developing culturally relevant classroom assessments. Applied Measurement in Education, 36(3), 286–303. https://doi.org/10.1080/08957347.2023.2214652

OECD. (n.d.). Education & Skills Online Assessment. https://www.oecd.org/skills/ESonline-assessment/abouteducationskillsonline/

OECD. (2015). Skills for social progress: The power of social and emotional skills. OECD Publishing. https://doi.org/10.1787/9789264226159-en

OECD. (2019). An OECD learning framework 2030. In G. Bast, E. G. Carayannis, & D. F. J. Campbell (Eds.), The future of education and labor. Arts, research, innovation and society (pp. 23–35). Springer. https://doi.org/10.1007/978-3-030-26068-2_3

OECD. (2021). AI and the future of skills: Volume 1. Capabilities and assessments. OECD Publishing. https://doi.org/10.1787/5ee71f34-en.

OECD. (2022a). Building the future of education. OECD Publishing. https://web-archive.oecd.org/2022-11-30/618066-future-of-education-brochure.pdf

OECD. (2022b). PISA 2022 results. https://www.oecd.org/publication/pisa-2022-results#pisa2022results

OECD. (2023). OECD skills outlook 2023: Skills for a resilient green and digital transition. OECD Publishing. https://doi.org/10.1787/27452f29-en

Oh, I.-S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. Journal of Applied Psychology, 96(4), 762–773. https://doi.org/10.1037/a0021832

O'Neil, H., Baker, E. L., Wainess, R., Chen, C., Mislevy, R., & Kyllonen, P. (2004). Final report on plan for the assessment and evaluation of individual and team proficiencies developed by the DARWARS Environments. Office of Naval Research; Defense Advanced Research Project Agency. https://apps.dtic.mil/sti/tr/pdf/ADA432802.pdf

OPM. (n.d.). Other assessment methods. OPM U.S. Office of Personnel Management. https://www.opm.gov/policy-data-oversight/assessment-and-selection/other-assessment-methods/

Ormerod, C. M., Malhorta, A.,&Jafari, A. (2021). Automated essay scoring using efficient transformer-based language models. PsyArXiv. https://arxiv.org/pdf/2102.13136.pdf

Ortner, T.M.,&Proyer, R. T. (2015). Objective personality tests. In T.M. Ortner&F. J. R. van de Vijver (Eds.), Behavior-based assessment in psychology: Going beyond self-report in the personality, affective, motivation, and social domains (pp. 133–149). Hogrefe.

Ortner, T.M., Proyer, R. T., & Kubinger, K. D. (2006). Theorie und praxis objektiver person-lichkeitstests [Theory and practice of objective personality tests]. Verlag Hans Huber.

Ostini, R., & Nering, M. L. (2006). Polytomous item response theory models (No. 144). Sage.

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(2), 187–207. https://doi.org/10.1037/0021-9010.89.2.187

Panadero, E. (2023). Toward a paradigm shift in feedback research: Five further steps influenced by self-regulated learning theory. Educational Psychologist, 58(3), 193–204. https://doi.org/10.1080/00461520.2023.2223642

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. Educational Research Review, 35, Article 100416. https://doi.org/10.1016/j.edurev.2021.100416

Panthier, C., & Gatinel, D. (2023). Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: A novel approach to medical knowledge assessment. Journal Français d'Ophtalmologie, 46(7), 706–711. https://doi.org/10.1016/j.jfo.2023.05.006

Patrick, S. (2021). Transforming learning through competency-based education. State Education Standard, 21(2), 23–29.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braubn, D. N. Jackson, & D. E. Wiley (Eds.), The role of constructs in psychological and educational measurement (pp. 49–69). Erlbaum.

Phelps, R. P. (2019). Test frequency, stakes, and feedback in student achievement: A meta-analysis. Evaluation Review, 43(3–4), 111–151. https://doi.org/10.1177/0193841X19865628

Poropat, A. E. (2014). A meta-analysis of adult-rated child personality and academic performance in primary education. British Journal of Educational Psychology, 84(2), 239–252. https://doi.org/10.1111/bjep.12019

Posso, A. (2016). Internet usage and educational outcomes among 15-year old Australian students. International Journal of Communication, 10, 3851–3876. https://ijoc.org/index.php/ijoc/article/view

Powers, D. E., & Fowles, M. E. (1997). The personal statement as an indicator of writing skill: A cautionary note. Educational Assessment, 4(1), 75–87. https://doi.org/10.1207/s15326977ea0401_3

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., & Black, A. W. (2018). Style transfer through back-translation. In I. Gurevych & Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Volume 1. Long Papers (pp. 866–876). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1080

Qian, Y., Tao, J., Suendermann-Oeft, D., Evanini, K., Ivanov, A. V., & Ramanarayanan, V. (2018a). Computer-implemented systems and methods for speaker recognition using a neural network (U.S. Patent 10,008,209). U.S. Patent Office and Trademark Office. https://ppubs.uspto.gov/pubwebapp/external.html?q=(10008209).pn.&db=USPAT&type=ids

Qian, Y., Tao, J., Suendermann-Oeft, D., Evanini, K., Ivanov, A. V., & Ramanarayanan, V. (2018b). Noise and metadata sensitive bottleneck features for improving speaker recognition with non-native speech input. In Proceedings of INTERSPEECH 2016: 17th Annual Conference of the International Speech Communication Association (pp. 3648–3652). https://doi.org/10.21437/Interspeech.2016-548

RAND. (2020). RAND education assessment finder. https://www.rand.org/education-and-labor/projects/assessments/tool.html

Randall, J. (2023). It ain't near 'bout fair: Re-envisioning the bias and sensitivity review process from a justice-oriented antiracist perspective. Educational Assessment, 28(2), 68–82. https://doi.org/10.1080/10627197.2023.2223924

Rees, A., (2021, December 27). The history of predicting the future. Wired. https://www.wired.com/story/history-predicting-future/

Rios, J. A., Ling, G., Pugh, R., Becker, D., & Bacall, A. (2020). Identifying critical 21st-century skills for workplace success: A content analysis of job advertisements. Educational Researcher, 49(2), 80–89. https://doi.org/10.3102/0013189X19890600

Roediger III, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. Journal of Experimental Psychology: Applied, 17(4), 382–395. https://doi.org/10.1037/a0026252

Roll, I., & Barhak-Rabinowitz, M. (2023). Measuring self-regulated learning using feedback and resources. In N. Foster & M. Piacentini (Eds.), Innovating assessments to measure and support complex skills. OECD Publishing. https://doi.org/10.1787/c93ac64e-en

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. Psychological Bulletin, 140(6), 1432–1463. https://doi.org/10.1037/a0037559

Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. Applied. Measurement in Education, 31(3), 191–214. https://doi.org/10.1080/08957347.2018.1464448

Rupp, A. A., Templin, J., & Henson, R. A. (2010). Diagnostic measurement: Theory, methods, and applications. Guilford Press.

Salgado, J. F., & Moscoso, S. (2019). Meta-analysis of interrater reliability of supervisory performance ratings: Effects of appraisal purpose, scale type, and range restriction. Frontiers in Psychology, 10, Article 2281. https://doi.org/10.3389/fpsyg.2019.02281

Salgado, J. F., & Tauriz, G. (2014). The Five-Factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. European Journal of Work and Organizational Psychology, 23(1), 3–30. https://doi.org/10.1080/1359432X2012.716198

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: a framework for constructing "intermediate constraint" questions and tasks for technology platforms. The Journal of Technology, Learning and Assessment, 4(6). https://ejournals.bc.edu/index.php/jtla/article/view/1653

Scalise, K., Malcom, C. & Kaylor, E. (2023). A tale of two worlds: Machine learning approaches at the intersection with educational measurement. In N. Foster&M. Piacentini (Eds.), Innovating assessments to measure and support complex skills (pp. 229–237). OECD Publishing. https://doi.org/10.1787/d01eb8a4-en

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), Educational measurement (4th ed.; pp. 307–353). American Council on Education; Praeger.

Schmill, S. (2022, March 28). We are reinstating our SAT/ACT requirement for future admissions cycles in order to help us continue to build a diverse and talented MIT. MIT Admissions. https://mitadmissions.org/blogs/entry/we-are-reinstating-our-sat-act-requirement-for-future-admissions-cycles/#annotation-10

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. Journal of Applied Psychology, 94(6), 1479–1497. https://doi.org/10.1037/a0016810

Schrum, L., & Levin, B. B. (2013). Leadership for twenty-first-century schools and student achievement: Lessons learned from three exemplary cases. International Journal of Leadership in Education, 16(4), 379–398. https://doi.org/10.1080/13603124.2013.767380

Schwartz, D. L., Tsang, J. M., & Blair, K. P. (2016). The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them. W.W. Norton & Company.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. Management Science, 58(8), 1438–1457. https://doi.org/10.1287/mnsc.1110.1509

Segall, D. O. (1996). Multidimensional adaptive testing. Psychometrika, 61(2), 331–354. https://doi.org/10.1007/BF02294343

Shafer, G. W., Viskupic, K., & Egger, A. E. (2023). Critical workforce skills for bachelor-level geoscientists: An analysis of geoscience job advertisements. Geosphere, 19(2), 628–644. https://doi.org/10.1130/GES02581.1

Shen, T., Lei, T., Barzilay, R.,&Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. In I. Gurevych, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in neural information processing systems 30 (NIPS 2017) (pp. 1–12). Curran Associates. https://papers.nips.cc/paper_files/paper/2017/file/2d2c8394e31101a261
Paper.pdf

Shepard, L. A. (2017). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), The future of assessment (pp. 279–303). Routledge. https://doi.org/10.4324/9781315086545-12

Shermis, M. D., & Burstein, J. (Eds.). (2013). Handbook of automated essay evaluation: Current applications and new directions. Routledge. https://doi.org/10.4324/9780203122761

Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153–189. https://doi.org/10.3102/0034654307313795

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. Adaptive technologies for Training and Education, 7(27), 1–35. https://doi.org/10.1017/CBO9781139049580.004

Sinatra, A. M., Robinson, R. L., Goldberg, B., & Goodwin, G. (2023). Impact of engaging with intelligent tutoring system lessons prior to class start. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67(1), 2262–2266. https://doi.org/10.1177/21695067231192709

Sinharay, S. (2023). Statistical methods for detection of test fraud on educational assessments. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.) International encyclopedia of education (4th ed., pp. 298–307). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10030-2

Sinharay, S., & Johnson, M. S. (2013). Statistical modeling of automatically generated items. In M. J. Gier & T. M. Haladyna (Eds.), Automatic item generation (pp. 183–195). Routledge.

Sinharay, S., & Johnson, M. S. (2023). Computation and accuracy evaluation of comparable scores on culturally responsive assessments. Journal of Educational Measurement, 61(1), 5-46. https://doi.org/10.1111/jedm.12381

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. Educational Measurement: Issues and Practice, 39(3), 100–105. https://doi.org/10.1111/emip.12377

Slavich, G. (2019). Stressnology: the primitive (and problematic) study of life stress exposure and pressing need for better measurement. Brain Behavior and Immunity, 75, 3–5. https://doi.org/10.1016/j.bbi.2018.08.011

Society for Industrial Organizational Psychology. (2018). Principles for the validation and use of personnel selection procedures (5th ed.). https://www.apa.org/ed/accreditation/personnel-selection-procedures.pdf

Soland, J.,&Kuhfeld, M. (2021). Do response styles affect estimates of growth on social-emotional constructs? Evidence from four years of longitudinal survey scores. Multivariate Behavioral Research, 56(6), 853–873. https://doi.org/10.1080/00273171.2020.1778440

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. Frontiers in Education, 4, Article 2019.00043. https://doi.org/10.3389/feduc.2019.00043

Solano-Flores, G. (2023). How serious are we about fairness in testing and how far are we willing to go? A response to Randall and Bennett with reflections about the Standards for Educational and Psychological Testing. Educational Assessment, 28(2), 105–117. https://doi.org/10.1080/10627197.2023.2226388

Soto, C. J., Napolitano, C. M., Sewell, M. N., Yoon, H. J., & Roberts, B. W. (2022). An integrative framework for conceptualizing and assessing social, emotional, and behavioral skills: The BESSI. Journal of Personality and Social Psychology, 123(1), 192–222. https://doi.org/10.1037/pspp0000401

Sottilare, R. A., Baker, R. S., Graesser, A. C., & Lester, J. (2018). Special issue on the generalized intelligent framework for tutoring (GIFT): Creating a stable and flexible platform for innovations in AIED research. International Journal of Artificial Intelligence and Education, 28(1), 139–151. https://doi.org/10.1007/s40593-017-0149-9

Sparks, J. R., Lehman, B., & Zapata-Rivera, D. (2024). Caring assessments: Challenges and opportunities. Frontiers in Education, 9, Article 1216481. https://doi.org/10.3389/feduc.2024.1216481

Stankov, L., Kleitman, S., & Jackson, S. A. (2015). Measures of the trait of confidence. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), Measures of personality and social psychological constructs (pp. 158–189). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-386915-9.00007-3

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. Journal of Educational Psychology, 105(4), 970–987. https://doi.org/10.1037/a0032447

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., Snook, S. A., & Grigorenko, E. L. (2000). Practical intelligence in everyday life. Cambridge University Press.

Stecher, B. M., & Hamilton, L. S. (2014). Measuring hard-to-measure student competencies: A research and development plan (Research Report No. RR-863-WFHF). RAND Corporation. https://doi.org/10.7249/RR863

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17(3), 277–292. https://doi.org/10.1177/014662169301700308

Stowe, K., Ghosh, D., & Zhao, M. (2022). Controlled language generation for language learning items. arXiv. https://doi.org/10.48550/arXiv.2211.15731

Straub, L. M., Lin, E., Tremonte-Freydefont, L., & Schmid, P. C. (2023). Individuals' power determines how they respond to positive versus negative performance feedback. European Journal of Social Psychology, 53(7), 1402–1420. https://doi.org/10.1002/ejsp.2985

Su, R., Tay, L., Liao, H.-Y., Zhang, Q., & Rounds, J. (2019). Toward a dimensional model of vocational interests. Journal of Applied Psychology, 104(5), 690–714. https://doi.org/10.1037/apl0000373

Tang, R., Chuang, Y.-N.,&Hu, X. (2023). The science of detecting LLM-generated texts. arXiv. https://doi.org/10.48550/arXiv.2303.07205

Tang, Z., & Kirman, B. (2023). Exploring curiosity in games: A framework and questionnaire study of player perspectives. International Journal of Human-Computer Interaction. Advance online publication. https://doi.org/10.1080/10447318.2024.2325171

Tannenbaum, R. J., & Kane, M. T. (2019). Stakes in testing: Not a simple dichotomy but a profile of consequences that guides needed evidence of measurement quality (Research Report No. RR-19-19). ETS. https://doi.org/10.1002/ets2.12255

Tenison, C., & Sparks, J. R. (2023). Combining cognitive theory and data driven approaches to examine students' search strategies in simulated digital environments. Large-Scale Assessments in Education, 11, Article 28. https://doi.org/10.1186/s40536-023-00164-w

Turchin, D. (Host). (2023, March 6). Andi Mann, Sageable CEO and AIOps pioneer, discusses enterprise AI wins and the impact of automation on jobs [Audio podcast episode]. In AI and the Future of Work. Apple Podcasts. https://podcasts.apple.com/us/podcast/andi-mann-sageable-ceo-and-aiops-pioneer-discusses/id1476885647?i=1000602978601

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. Annual Review of Clinical Psychology, 9, 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510

U.S. Congress, Office of Technology Assessment. (1992). Testing in American schools: Asking the right questions (Report No. OTA-SET-519). U.S. Government Printing Office.

U.S. Office of Personnel Management. (n.d.). Situational judgment tests. https://www.opm.gov/policy-data-oversight/assessment-and-selection/other-assessment-methods/situational-judgment-tests/

van der Linden, W. J. (2005). Linear models for optimal test design. Springer. https://doi.org/10.1007/0-387-29054-0

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). Elements of adaptive testing. Springer. https://doi.org/10.1007/978-0-387-85461-8

van de Vijver, F. J. R.,& He, J. (2016). Bias assessment and prevention in noncognitive outcome measures in context assessments. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), Assessing contexts of learning: International perspectives (pp. 229–253). Springer. https://doi.org/10.1007/978-3-319-45357-6_9

van de Vijver, F., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. H. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), Adapting educational and psychological tests for cross-cultural assessment. Taylor & Francis. https://doi.org/10.4324/9781410611758

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? Cognitive Science, 31(1), 3–62. https://doi.org/10.1080/03640210709336984

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. Psychological Test and Assessment Modeling, 52(1), 8–28.

von Davier, M., Tyack, L., & Khorramdel, L. (2023). Scoring graphical responses in TIMSS 2019 using artificial neural networks. Educational and Psychological Measurement, 83(3), 556–585. https://doi.org/10.1177/00131644221098021

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, Article 106189. https://doi.org/10.1016/j.chb.2019.106189

Wainer, H. (1987). The first four millennia of mental testing: From ancient China to the computer age (Research Report No. RR-87-34). ETS. https://doi.org/10.1002/j.2330-8516.1987.tb00238.x

Wainer, H., & Thissen, D. (2001). True score theory: The traditional method. In D. Thissen & H. Wainer (Eds.), Test scoring (pp. 35–84). Routledge. https://doi.org/10.4324/9781410604729

Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D.,&Gholson, M. (2023). Culturally responsive assessment: provisional principles (Research Report No. RR-23-11). ETS. https://doi.org/10.1002/ets2.12374

Walkington, C., & Bernacki, M. L. (2020). Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. Journal of Research on Technology in Education, 52(3), 235–252. https://doi.org/10.1080/15391523.2020.1747757

Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., Yin, Y., Huang, Z.,&Wang, S. (2020). Neural cognitive diagnosis for intelligent education systems. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), 6153–6161. https://doi.org/10.1609/aaai.v34i04.6080

Wang, J., Jou, M., Lv, Y., & Huang, C. C. (2018). An investigation on teaching performances of model-based flipping classroom for physics supported by modern teaching technologies. Computers in Human Behavior, 84, 36–48. https://doi.org/10.1016/j.chb.2018.02.018

Weinberger, C. J. (2014). The increasing complementarity between cognitive and social skills. Review of Economics and Statistics, 96(5), 849–861. https://doi.org/10.1162/REST_a_00449

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. Applied Psychological Measurement, 41(2), 115–129. https://doi.org/10.1177/0146621616676791

Weiss, S., Wilhelm, O., & Kyllonen, P. (2021). An improved taxonomy of creativity measures based on salient task attributes. Psychology of Aesthetics, Creativity, and the Arts. Advance online publication. https://doi.org/10.1037/aca0000434

West, M., Pier, L., Fricke, H., Hough, H. J., Loeb, S., Meyer, R. H., & Rice, A. B. (2018). Trends in student social-emotional learning: evidence from the CORE districts (Working paper). Policy Analysis for California Education. https://edpolicyinca.org/publications/trends-student-social-emotional-learning

Williamson, D.M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wilkie, D. (2023, December 21). Employers say students aren't learning soft skills in college. https://www.shrm.org/topics-tools/news/employee-relations/employers-say-students-arent-learning-soft-skills-college

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. Educational Measurement: Issues and Practice, 36(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. Educational Assessment, 10(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. Frontiers in Psychology, 10, Article 3087. https://doi.org/10.3389/fpsyg.2019.03087

Wolcott, M. D., Lobczowski, N. G., Zeeman, J. M., & McLaughlin, J. E. (2020). Situational judgment test validity: an exploratory model of the participant response process using cognitive and think-aloud interviews. BMC Medical Education, 20, Article 506. https://doi.org/10.1186/s12909-020-02410-z

World Economic Forum. (2021). Building a common language for skills at work: A global taxonomy. https://www3.weforum.org/docs/WEF_Skills_Taxonomy_2021.pdf

World Economic Forum. (2022). Catalysing Education 4.0: Investing in the future of learning for a human-centric recovery. https://www3.weforum.org/docs/WEF_Catalysing_Education_4.0_2022.pdf

World Economic Forum. (2023). Defining Education 4.0: A taxonomy for the future of learning. https://www3.weforum.org/docs/WEF_Defining_Education_4.0_2023.pdf

Xuan, Q., Cheung, A., & Sun, D. (2022). The effectiveness of formative assessment for enhancing reading achievement in K-12 classrooms: A meta-analysis. Frontiers in Psychology, 13, Article 990196. https://doi.org/10.3389/fpsyg.2022.990196

Yang, Z., Hu, Z., Dyer, C., Xing, E. P., & Berg-Kirkpatrick, T. (2018). Unsupervised text style transfer using language models as discriminators. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), Advances in neural information processing systems 31 (NeurIPS 2018) (pp. 7287–7289). Curran Associates. https://papers.neurips.cc/paper_files/paper/2018/hash/398475c83b47075e8897a083e97eb9f0-Abstract.html

Yeung, C. (2019). Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. arXiv. https://doi.org/10.48550/arXiv.1904.11738.

Yeung, C.-K., & Yeung, D.-Y. (2019). Incorporating features learned by an enhanced deep knowledge tracing model for STEM/non-STEM job prediction. International Journal of Artificial Intelligence and Education, 29, 317–341. https://doi.org/10.1007/s40593-019-00175-1

Youyou, W., Kosinski, M.,&Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Zapata-Rivera, D., & Forsyth, C. M. (2022). Learner modeling in conversation-based assessment. In R. A. Sottilare, & J. Schwarz (Eds.), Adaptive instructional systems: International Conference on Human-Computer Interaction. HCII 2022 (pp. 73–83). Springer. https://doi.org/10.1007/978-3-031-05887-5_6

Zapata-Rivera, D., Graesser, A. C., Kay, J., Hu, X.,&Ososky, S. J. (2020). Visualization implications for the validity of intelligent tutoring systems. In A. M. Sinatra, A. C. Graesser, X. Hu, B. Goldberg, & A. J. Hampton (Eds.), Design recommendations for intelligent tutoring systems: Volume 8. Data visualization (pp. 61-68). US Army Combat Capabilities Development Command - Soldier Center.

Zapata-Rivera, D., & Hu, X. (2022). Assessment in intelligent tutoring systems SWOT analysis. In A. M. Sinatra, A. C. Graesser, X. Hu, G. Goodwin,&V. Rus (Eds.), Design recommendations for intelligent tutoring systems: Vol. 10. Strengths, weaknesses, opportunities and threats (SWOT) analysis of intelligent tutoring systems (pp. 83–90). US Army Combat Capabilities Development Command – Soldier Center. https://gifttutoring.org/attachments/download/4751/Vol%2010_DesignRecommendationsforITSs_SWOTAnal

Zapata-Rivera, D., Lehman, B., & Sparks, J. R. (2020). Learner modeling in the context of caring assessments. In R. A. Sottilare & J. Schwarz (Eds.), Adaptive instructional systems: Second International Conference (AIS) 2020 (pp. 422–431). Springer.

Zhan, J., Her, Y. W., Hu, T., & Du, C. (2018). Integrating Data Analytics into the Undergraduate Accounting Curriculum. Business Education Innovation Journal, 10(2), 169–178. http://www.beijournal.com/images/V10N2_draft_5.pdf

Zhang, Z. (2012). Microsoft Kinect sensor and its effect. IEEE multimedia, 19(2), 4–10. https://doi.org/10.1109/MMUL.2012.24

Zu, J., & Choi, I. (2023a, April 12–15). Utilizing deep language models to predict item difficulty of language proficiency tests [Paper presentation]. The annual meeting of National Council on Measurement in Education, Chicago, IL, United States.

Zu, J., & Choi, I. (2023b, July 25–28). Predicting the psychometric properties of automatically generated items [Paper presentation]. International Meeting of the Psychometric Society, College Park, MD, United States.

Zu, J., Choi, I., & Hao, J. (2023). Automated distractor generation for fill-in-the-blank items using a prompt-based learning approach. Psychological Testing and Assessment Modeling, 65(2), 55–75.

Zu, J.,&Kyllonen, P. C. (2020). Nominal response model is useful for scoring multiple-choice situational judgment tests. Organizational Research Methods, 23(2), 342–366. https://doi.org/10.1177/1094428118812669