4 Occupational tests

Mila Staneva, OECD

Britta Rüschoff, FOM University of Applied Sciences for Economics and Management

Phillip L. Ackerman, Georgia Institute of Technology

This chapter describes performance tests on occupational tasks stemming from occupation certification and licensure examinations. It discusses use of such examination tasks for collecting expert judgement on artificial intelligence (AI) and robotics performance. The chapter describes 13 example tasks from six occupations selected for an explanatory assessment of AI and robotics. The tasks were chosen from final examinations in German vocational education and training, as well as certification and licensure exams used in the United States. The chapter describes the development and administration of such tests, their types and formats, as well as procedures to ensure their content validity. It concludes with discussing methodological steps towards a comprehensive and robust approach for studying the capabilities of AI and robotics and their impact on occupations.

Understanding the implications of evolving technologies for education and employment requires an evaluation of artificial intelligence (AI) and robotics across a wide range of skills used in the workplace. Next to key cognitive skills, such as literacy and numeracy, the workplace involves various occupation-specific technical skills and domain-specific professional knowledge. Accordingly, a battery of different instruments for measuring AI and robotics capabilities is needed.

As a complement to the education tests discussed in Chapter 3, this chapter explores the use of complex occupational tasks for collecting expert ratings on AI and robotics capabilities. These tests stem from licensing and certification examinations for different occupations. They include typical (hands-on) tasks in the occupation, such as a product designer creating a design for a new container lid, or a cosmetologist performing a manicure. Experts' ratings of AI and robotics' performance on such tasks can provide valuable insights into the readiness of these technologies for real-world applications and their potential to replace or support workers in their jobs.

This chapter discusses the usefulness of complex occupational tasks for evaluating AI and robotics. Tasks from two sources are presented – final examinations in German vocational education and training (VET) as well as certification and licensure exams used in the United States. The chapter describes the development and administration of such tests, their types and formats, and procedures to ensure content validity. It then outlines 13 tasks from six occupations selected for an exploratory evaluation of AI and robotics' performance using expert judgement. The chapter concludes with discussing further steps towards an approach for assessing AI and robotics capabilities across a wide range of occupational tasks and analysing how progress in capabilities may change these tasks.

Rationale for collecting expert judgement on AI with complex occupational tasks

Like other human tests used to evaluate AI in this project, examinations that certify workers for specific occupations have a variety of advantages. They offer computer experts standardised and objective evaluation criteria for rating AI and robotics' performance. This allows for consistent ratings across different expert groups and across time. They provide precise, contextualised and granular descriptions of the tasks. This allows experts to make exact judgements of AI's potential performance on the task (and parts of the task) and, thus, improves reliability across experts. Moreover, occupation certification and licensure tests provide a way to compare AI and robotics performance to human performance. Where data on the performance of human test takers are missing, the passing score on the exam can indicate whether machines satisfy the minimum skills requirements that workers must fulfil to enter occupations.

An additional advantage of certification and licensure examinations is related to the use of real-world tasks and scenarios that are typical for occupations. These tasks are action-oriented and practical, drawing on observable task-related behaviour (e.g. crafting a certain product). The practical relevance is provided in several ways. For example, test items are selected on the basis of careful job analysis that determines the most important and most frequently applied tasks in occupations (Johnston et al., 2014[1]). People working in the profession, so-called "subject matter experts", and/or industry representatives are often involved in the test design. Examinations are aligned to the framework curricula of training programmes, which, for their part, mirror industry standards and best practices (Rüschoff, 2019[2]).

This practical orientation of occupation certification and licensure tests distinguishes them from other types of tests that rely on broad, underlying characteristics of applicants to predict job performance. The latter tests seek to assess abstract constructs, such as general intelligence, broad content abilities (e.g. verbal, spatial, numerical abilities) or narrower abilities (e.g. perceptual speed, psychomotor abilities), based on evidence that these traits manifest in various behaviours, including how one performs job-related tasks. For example, individuals who score high on a general mental ability test are more likely to successfully perform tasks involving complex problem-solving in real-life situations. The tasks included in the ability test can be thus indicative for an observable behaviour of interest. However, these tasks are indicators of an

abstract, underlying psychological construct rather than direct demonstrations of the behaviour. They are meaningful insofar as they are correlated with the behaviour of interest.

By contrast, the occupational tasks included in certification and licensure examinations have considerable meaning on their own. They assess concrete professional behaviours with immediate relevance to exercise of the profession. The resulting measures rely far less on theoretical assumptions regarding underlying psychological constructs. This poses an advantage for assessing AI and robotics since a theoretical link between concrete performance on a task and broad underlying abilities cannot be assumed for machines. In other words, high performance on a task cannot be attributed to a general ability that would enable high performance on another, different task (OECD, 2023[3]).

As shown in Chapter 3, the use of tests targeting broad underlying foundation skills, such as literacy and numeracy, posed challenges to the assessment. Experts diverged in their ratings of AI performance on these tests because they were uncertain how general the computer capabilities being assessed are supposed to be. They argued that the capacity of systems to solve one test task does not presuppose high performance on other task types and formats. To make precise judgements, the experts thus needed clarification on the generality of the underlying capabilities being evaluated. However, defining generality is not trivial. It requires a specification of all tasks that a system is supposed to master, within and beyond the test.

Testing AI and robotics on concrete occupational tasks should mitigate this problem. The measures of this type of task performance are also narrow in their generalisability. The reason is that occupational tasks are complex, involving multiple actions and requiring different capabilities, which makes it hard to attribute success on one task to other tasks or unknown contexts. Still, performance tests from occupation entry examinations would show whether a machine can or cannot complete typical tasks in an occupation.

Occupational tasks from certification and licensure examinations

Many occupations require passing an exam that establishes whether candidates demonstrate the requisite knowledge, skills and abilities to engage in practice. Additional minimum entry requirements are often in place, such as qualifications, experience or medical record. These occupational entry regulations serve the purpose of protecting the public from unqualified practitioners and ensuring good quality services and products through standardising the skills of their providers (Koumenta and Pagliero, 2017[4]). Such considerations are especially strong for occupations that are of particular interest for the public good.

Occupational entry regulations can take different forms. Licensure is the strictest form of regulation. It grants those who can demonstrate the specified level of competence the legal right to exercise protected activities. Persons without a licence cannot practice the occupation. By contrast, certification provides a legally protected title that indicates a minimum competence for an occupation. Those who do not hold the certificate are not legally restricted from carrying out tasks covered by the occupation. While licensing is overseen by government or state authorities (or appointed regulators), certification programs can also be developed by professional associations, chambers of industry or other membership organisations.

Occupation licensure and certification practices vary across countries and occupations. In the European Union, 43% of workers held a certificate or a licence in 2015 (Koumenta and Pagliero, 2017[4]). The proportion of licensed workers was highest in Germany, at 33%, and lowest in Denmark, at 14%. The proportion of certified workers varied between 36% in Germany and 9% in Finland (Koumenta and Pagliero, 2017[4]). In the United States, 28% of the workforce was licensed or certified in 2013, with big differences among the states (Kleiner and Vorotnikov, 2017[5]). In general, occupational licensing and certification are more typical for teaching professionals, health and social workers, and plant and machine operators, and is less common for managers, in wholesale or retail services, agriculture or elementary occupations (Koumenta and Pagliero, 2017[4]). However, the same occupation can be subject to very

different entry regulations in different jurisdictions. In the United States, for example, fewer than 60 occupations were regulated in all 50 states. Meanwhile, more than 1 000 occupations were regulated in at least one state (Kleiner and Vorotnikov, 2017[5]).

To account for possible country differences in examination practices, this study uses occupational tests from two countries – Germany and the United States.

German VET assessments

Germany has 324 vocational occupations that are state-recognised under the Vocational Training Act (BBiG) or the Crafts Code (HwO) (BIBB, 2022[6]). Other state-regulated occupations, such as in the medical field, are covered in special laws (e.g. Nursing Act, Geriatric Care Act). The dual VET system provides entry into most vocational occupations. This study refers primarily to dual VET since it is the most common training model that qualifies workers for occupations in Germany.

In dual VET, apprentices acquire theoretical knowledge by attending a vocational school and receive practical training by working in a company. Apprentices sign a contract with the company and receive renumeration for their work. In-company training is regulated by the Vocational Training Act and by the training regulations of the occupations (*Ausbildungsordnungen*). School-based vocational education is regulated by framework curricula (*Rahmenlehrpläne*). The training regulations and the framework curricula provide national standards regarding training content, training facilities, trainers and examinations (Cedefop, 2020[7]).

At the end of their training, apprentices complete a final exam to obtain a certificate. Final examinations in (dual) VET are regulated by the Vocational Training Act (BBiG, §37 – §50) or the Crafts Code (HwO, §31 – §40a). They are organised by the respective chambers for each occupation. The chambers appoint examination boards and conduct the examinations, which are developed in accordance with the relevant regulatory instruments – training regulations and the framework curriculum (OECD, 2021[8]).

Examinations are aligned with the curricula to reflect all relevant domains in the occupation. Examples of professional behaviour in the respective occupation are provided for each examination domain. These examples form the basis for the examination tasks. The tasks are commonly classified by content and the competences they aim to assess (Badura, 2015[9]).

Different test developers follow different competence models when developing tasks. For example, AkA (*Aufgabenstelle für kaufmännische Abschluss- und Zwischenprüfungen*)[1], which develops examinations for commercial occupations, classifies tasks according to the content domains planning, execution and evaluation of results and according to whether tasks assess knowledge or skills and abilities (Badura, 2015[9]). The classifications used are not always aligned with international skills taxonomies commonly considered in research and policy.

Examination development offices usually develop the tasks. For example, PAL (*Prüfungsaufgaben- und Lehrmittelentwicklungsstelle*)[2] develops exams for industrial and technical professions, ZPA (*Zentralstelle für Prüfungsaufgaben*)[3] or AkA cover commercial professions, and the ZFA (*Zentral-Fachausschuss Berufsbildung Druck und Medien*)[4] provides examinations in the field of printing and media. However, other entities may also develop tasks. Examples include the examination board appointed by the Chamber of Commerce and Industry (IHK) or a task development committee appointed by the respective chambers.

The tasks are developed in close co-operation with the industry to ensure their content validity i.e. that the task content fully represents the requirements and content of the occupation. Another instrument for establishing content validity are examination catalogues or grids. An examination grid indicates the proportional distribution of the tasks across examination domains. It is based on the specifications in the regulatory instruments. The grid is intended to ensure the examination covers all content domains relevant in an occupation and to give domains a correct weighting. In addition to the validity of the examinations,

the development offices commonly keep a record of the reliability (i.e. whether a test produces similar results under consistent conditions) and discriminatory power of the tests (i.e. whether a test can distinguish between two or more groups being assessed).

US licensing and certification exams

In the United States, occupational credentialing and licensure are largely decentralised. State governments generally enact the laws regulating occupational licensing. Some states embed these requirements directly in the statute authorising creation of the licence. Other states authorise their agencies or state-sponsored independent boards to develop licensure requirements. Often, occupation entry requirements combine both – statute and regulations set by a designated agency or board (NCSL, 2022[10]).

As a result, there are significant differences in licensing requirements across the states (NCSL, 2022[10]). For example, the State of Georgia has 180 different occupational licences. Many are different "levels" or types of the same occupation, such as separate certifications for nine different nurse licences. In comparison, the State of California has 357 different occupational licences (US Department of Labor, 2021[11]). In some instances, reciprocity agreements make it easier for licensees in one state to be licensed in another.

The states typically delegate implementation of occupational entry regulations to professional associations. The latter usually form one or more intermediary agencies to assume responsibility for development and validation of examinations (e.g. National Council of Architectural Registration Boards, American Board of Dental Examiners). In other instances, agencies or organisations that award credentials develop examinations for credentials on their own (e.g. National Commission on Certification of Physician Assistants (Buckendahl, 2017[12]). In addition, third-party auditors such as the National Council of Certifying Agencies and the American National Standards Institute have developed formal standards for evaluating credentialing programmes (Johnston et al., 2014[1]).

Test development for certification or licensure is typically a function of formal or informal job analyses, the engagement of subject matter experts, and attention to a body of reference materials for formulating test questions and determining the scope of the examinations. Job analysis involves different methods to identify the mainstream activities of the profession. In addition, it collects information about the knowledge, skills and abilities (KSAs) needed for performing these activities. This information is typically collected from surveys of subject matter experts. Other sources of information include course outlines, laws, textbooks or other curricular materials. The apparent goal of these design characteristics is to improve the content validity of the test rather than construct validity (i.e. Does the test measure the construct it intends to measure?) or criterion-related validity (i.e. Do test results correlate with results from other tests measuring the construct?).

The results of the job analysis are used to develop a so-called test blueprint. This document serves as the basis for developing concrete examination tasks by specifying the most important characteristics of the test. These are the behaviours or KSAs to be assessed, but also other features, such as the emphasis given to different domains, the item format or the difficulty of the test. In addition to providing item developers with direction, blueprints help ensure continuity in test content and difficulty over time and serve as the basis for item classification and revision. They can also inform educators and test takers of test content and assist them in their test-preparation efforts.

Formats of performance tests and grading

The tasks included in certification and licensure examinations can be written, oral or practical (hands-on procedural) demonstrations of knowledge and skills. Written tasks can have different formats (e.g. open-answer, multiple-choice, fill-in-the-blank questions). They are often applied, in the sense that they cover typical professional activities, such as writing a business letter in commercial professions or

documenting the manufacture of a product in technical professions. Examples for oral tasks are presentations, conversation simulations or discussing a completed assignment. Procedural demonstrations typically include crafting a product or carrying out a typical activity for the profession (BIBB, 2013[13]).

The grading of occupation examinations varies widely. Licensure examinations often use a binary pass/fail grading since they aim at determining whether the examinee has a "minimum competence" for the profession. Different cut scores can determine whether an examinee passes or fails. Some examination procedures use absolute scores (e.g. 60% correct). Other examinations use "scaled scores". These are essentially norm-referenced measures where passing depends on the individual's percentile rank among other test takers. Classifying an examinee as qualified/non-qualified is thus determined partly by how others perform in the examination.

By contrast, certification exams usually use continuous-graded scales or categorical ratings, such as "novice, apprentice, journeyman and expert" (Hambrick and Hoffman, 2016[14]). This grading is more suitable for evaluating computer performance on occupational tasks and comparing it to human performance. However, even when examinations designate individual examinees as qualified or not qualified, there is an underlying continuous scale upon which individuals perform. Thus, the application of occupation tests for evaluating AI and robotics should not be limited to a dichotomous evaluation.

Selection of occupations and examination tasks

The project selected 13 example tasks from six occupations – amid a larger pool of identified occupations – to explore their use for assessing AI and robotics capabilities. The aim was to select diverse examination tasks representing some important elements of reasoning, language and sensory-motor capabilities. In addition, the tasks covered different occupations and working contexts and had different levels of complexity to explore how these different aspects relate to the collection of expert judgement.

Several considerations guided the selection of occupations:

First, occupations were sampled across the broad categories of the International Standard Classification of Occupations (ISCO) (ILO, 2012[15]). ISCO divides occupations into ten major groups, which are further divided into smaller subgroups. The categorisation of occupations into broad groups depends on the skill level and the education required for occupations. Occupations from five of the ten major groups were selected to cover different levels of the occupational hierarchy (see Table 4.1).

Table 4.1. Selected occupations

Occupation	ISCO occupational domain	ISCO 4-digit code	NACE Industry sector
Specialist in metal technology	7 Craft and related trades workers	7212 - 7215	C Manufacturing
Cosmetologist	5 Service and sales workers	5141	S Other service activities
Office management assistant	4 Clerical support workers	4110	N Administrative and support service activities
Dental assistant	3 Technicians and associate professionals	3251	Q Human health and social work activities
Nursing professional	2 Professionals	2221	Q Human health and social work activities
Technical product designer	2 Professionals	2163	M Professional, scientific and technical activities

Note: The International Standard Classification of Occupations (ISCO) has ten broad groups according to required skill level and qualification: 1 Managers; 2 Professionals; 3 Technicians and associate professionals; 4 Clerical support workers; 5 Service and sales workers; 6 Skilled agricultural, forestry and fishery workers; 7 Craft-related trades workers; 8 Plant and machine operators, and assemblers; 9 Elementary occupations; 0 Armed forces occupation (ILO, 2012[15]). The Statistical Classification of Economic Activities in the European Community (NACE) is an international taxonomy of industries that distinguishes 21 major industrial sectors (European Commission, 2008[16]).

Box 4.1. Direct assessment of large language models on written professional certification tests

A number of studies applied GPT (Generative Pre-trained Transformer) language models on written professional certification exams to study their content-specific capabilities. For example, Noever and Ciolino (2023[17]) assessed the performance of GPT-3 and GPT-3.5 on a test dataset containing 1 149 professional certifications. They showed that GPT-3 could solve more than 70% of the test questions on 39% of the exams. GPT-3.5 demonstrated better performance on many exams, particularly those for accountants, veterinarians, aviation inspectors, real estate appraisers, human resources professionals and financial planners.

Several studies evaluated the large language models in the domain of medicine. For example, Ali et al. (2023[18]) tested GPT-3.5 and GPT 4 on a neurosurgery written examination. GPT-3.5 scored 73.4% and GPT-4 scored 83.4% on the test, compared to an average of 73.7% of human tests takers. Antaki et al. (2023[19]) evaluated ChatGPT, based on GPT-3.5, in the domain of ophthalmology and found that the model had modest overall performance. By contrast, Lin et al. (2023[20]) evaluated both GPT-3.5 and GPT-4 on ophthalmology written examination and found that GPT-4 performance (76.9%) exceeds both the performance of its predecessor (63.1%) and human performance (72.6%). Nori et al. (2023[21]) tested GPT-4 on practice materials for the United States Medical Licensing Examination (USMLE). They show that the model exceeds the passing score of the test by 20 points and outperforms GPT-3.5 as well as models specifically developed for the medical field.

Other studies showed that GPT-4 passed US license exams in accounting (Eulerich et al., 2023[22]) and law (Katz et al., 2023[23]).

While these studies evaluate large language models on written exams, the exploratory study described in Chapter 5 asks experts to assess the state of the art in AI and robotics on mostly practical tasks from occupational examinations. As described in this chapter, the examination materials used in this study are diverse, testing capabilities such as vision, planning or dexterity.

Second, occupations from different industries were selected. Following the Statistical Classification of Economic Activities in the European Community, commonly referred to as NACE, five industries were covered: manufacturing; health; professional, scientific and technical activities; administrative and support service activities; and other service activities (European Commission, 2008[16]).

Different examination materials were retrieved for each selected occupation (see Table 4.2). Due to the large scope of the examinations, only parts of the exams or single tasks were used. The criteria for selecting these materials aimed again at diversifying the set of examination tasks in order to test different possibilities for assessing AI and robotics performance on the job. First, materials from both US and German examinations were selected. Second, both practical and written tasks were used. The written tasks contained both open (e.g. writing an e-mail) and closed (e.g. fill-in-the-blank task) questions. The practical tasks ranged from shorter tasks to comprehensive work assignments that took examinees several hours to complete. Some materials contained specific instructions, including a list of the instruments available to the examinees and supplementary materials, such as technical drawings or plans. Other tasks were described briefly. This aimed to test how much information on each task experts need to reliably judge AI and robotics potential performance. Finally, tasks were chosen to cover a wide range of skills, from sensory-motor skills to reasoning and language and use of domain-specific knowledge. (Noever and Ciolino, 2023[17])

Table 4.2. Selected occupational tasks

Occupation	Task content	Examples of skills required in the task	Task format and provider
Specialist in metal technology	Manufacture a functional assembly according to given specifications; assess and document whether the components of the product are dimensionally accurate in a measurement protocol.	Knowledge of tools and materials; technical knowledge of the process steps in manufacturing products; planning; ability to autonomously execute work orders according to technical instructions; subject-specific mathematical skills, e.g. reading units of measurement; general dexterity; general mathematical skills; spatial thinking/spatial imagination.	Practical task (DEU)[1]
	Use two pieces of sheet aluminium and filler rod to weld a Tee-joint in the horizontal position.	Ability to select and set up equipment correctly and safely; ability to select and use the right material (aluminium and filler rod); ability to weld according to specifications.	Practical task (US)[2]
Cosmetologist	Perform chemical waving.	Knowledge of chemical waving supplies; ability to wrap hair; ability to place rod correctly throughout entire section; ability to understand and follow instructions.	Practical task (US)[3]
	Perform a full manicure including a hand massage, remove excess cream from nails, and polish nails.	Knowledge of materials and tools; ability to use correct manicure techniques; ability to perform a hand massage; ability to understand and follow instructions; ability to clean workstation.	Practical task (US)[2]
Office management assistant	Prepare an evaluation of the complaints based on available data; find several files saved in different folders and use spreadsheets from these files according to given specifications.	ICT skills; general literacy skills; general mathematical skills; ability to read tables; ability to perform spreadsheet calculations (e.g. application of formulas); analytical ability to translate data into meaningfully visualised diagrams; ability to analyse business data and to communicate the results of these analyses appropriately.	Practical task (DEU)[4]
	Draft an e-mail to line manager explaining the results of the analysis of the complaints and proposing solutions.	ICT skills; general literacy skills; ability to write professional communication (e-mail); ability to formulate written communication (by e-mail) appropriate to the addressee; ability to handle and understand tables/data; ability to autonomously develop ideas for solutions based on data and professional knowledge; ability to communicate own ideas and proposed solutions in a comprehensible way (in writing).	Writing task (DEU)[4]
	Create a flyer using specifications provided.	ICT skills; general literacy skills; ability to understand and follow instructions; familiarity with formatting techniques and practices; ability to use and maintain office equipment (copier).	Practical task (US) [2]
Dental assistant	Indicate the different groups of teeth and their distribution in the deciduous and permanent dentition in the blank spaces of the table provided.	Knowledge of Latin dental terminology; know differences between deciduous and permanent dentition; general literacy skills.	Writing task (DEU)[5]
	Name two ways of performing a sensitivity test.	Knowledge of dental terminology; knowledge of dental exam procedures; ability to identify correct dental procedure based on a given situation and exam purpose.	Writing task (DEU)[5]
	Prepare instruments for autoclaving.	Knowledge of correct materials and equipment for autoclaving; knowledge of pre-cleaning, disinfection and sterilisation procedures according to federal guidelines; ability to apply the correct safety and sanitation procedures in preparing dental instruments; ability to understand and follow given instructions.	Practical task (US) [2]
Nursing professional	Transfer a Cerebral Vascular Accident patient with right-side paralysis from bed to wheelchair and back to bed.	Knowledge of necessary equipment; ability to identify the patient; ability to introduce and explain the procedure; ability to use equipment and aseptic techniques properly and safely; ability to identify appropriate patient positioning, transfers and body alignment; ability to position patient in a wheelchair and bed.	Practical task (US) [2]

Occupation	Task content	Examples of skills required in the task	Task format and provider
Technical product designer	Produce technical drawing proposals of modifications to a kitchen tool using Computer-Aided Design (CAD).	Ability to translate technical drawings and functional descriptions into work orders; planning skills; general literacy; analytical/mathematical ability to calculate and complete missing dimensions autonomously using known parameters; ability to make autonomous design decisions based on a functional description and known parameters; ability to produce different types of drawings (e.g. exploded view); basic technical/physical knowledge; ability to use CAD.	Practical task (DEU)[6]
	Create a 3D solid model using CAD.	Ability to produce precise technical drawings using CAD; general literacy; basic technical/physical knowledge; ability to make autonomous design decisions based on a functional description and known parameters; spatial imagination.	Practical task (US) [2]

Note: 1) PAL (2014[24]); 2) Materials come from test blueprints made available by NOCTI at https://www.nocti.org.; 3) https://nictesting.org 4) AkA (2021[25]); 5) Zahnärztekammer Niedersachsen (2021[26]); 6) PAL (2019[27]).

The way forward

The project selected 13 tasks from six occupations for an exploratory assessment of AI and robotics performance on work tasks. The selected tasks cover diverse occupations and skills, allowing to test assessment methodologies in different set-ups. However, the selection represents only an excerpt of the wide occupational space. A comprehensive assessment of AI and robotics performance in occupations would require a much larger effort. The project will need to assess AI and robotics on a larger set of tasks that represent the variety of skills and knowledge used in the workplace. It should also attempt to study how AI progress in these skills changes occupational tasks and occupations.

A first step towards a comprehensive AI assessment is the development of a systematic approach for sampling occupations and occupational tasks. This sampling approach should rely on clearly defined criteria for selecting occupations. It should aim at producing a feasibly sized sample of work tasks that sufficiently represents key dimensions of human performance at work (e.g. knowledge, skills, abilities, proficiency). In addition, it should adequately cover various work contexts, spanning the different occupational and industry domains, as well as tasks of varying complexity and generality.

As one sampling challenge, not all occupations are subject to entry examinations. As described above, occupation entry examinations are less common in elementary occupations, agriculture, managerial and sales jobs (Koumenta and Pagliero, 2017[4]). This may result in underrepresentation of these occupations in the assessment. To address this problem, the project will seek alternatives to entry examinations for unregulated occupations. One way would be to consult experts on the most common tasks in these occupations. Another way is to obtain examples of common tasks from occupational taxonomies, such as the Occupational Network (O*NET) database of the US Department of Labor (National Center for O*NET Development, n.d.[28]).

As another shortcoming, certification and licensing examinations are too focused on assessing professional skills. Consequently, they often neglect other types of skills that are equally relevant at the workplace, such as general cognitive skills or social skills. A systematic review on the methods of competence assessment in German VET showed that 60% of assessment instruments used in examinations targeted occupation-specific competences. However, only 24% assessed general competencies, such as writing, reading and mathematics. Another 9% focused on social competences such as communication skills (Rüschoff, 2019[2]). This highlights the need for a battery of different instruments to assess AI and robotics, including education tests (Chapter 3) and AI benchmark tests (Chapters 6-8), to capture a wider array of skills.

In its third phase, the project will aim at developing methods for studying how occupations may change in response to evolving AI and robotics. AI and robotics will not simply substitute workers in occupational tasks. In some tasks, machines will support workers by completing only parts of the work. This may result in additional tasks for the worker, such as monitoring and supervising the computer systems, thus changing skill requirements for the job. Moreover, the application of AI and robotics may completely change the task by reinventing its solutions or by altering the work environment to adapt it to the use of machines.

To study the implications of evolving AI and robotics for occupations, the project will develop a set of task descriptions to illustrate how different human tasks are likely to evolve as we begin to carry them out with AI support. The set of tasks would need to illustrate the full range of tasks that are carried out at work and in everyday life, including tasks with key cognitive, physical and social aspects. For each of these tasks, the project will develop feasible scenarios for the way the tasks will likely be carried out with AI support. The goal is to have a group of AI experts, job analysts and psychologists analyse each of the sampled tasks to determine which activities could be performed by an AI system and then propose ways for redesigning the current task to allow a human to work with the support of an AI system. This would make it possible to describe a transformed role for humans in each of the tasks. The analysis would be carried out for each of the sampled tasks, considering current AI performance levels for the different capabilities, as well as several scenarios for future performance levels that AI could plausibly achieve in the next 5-20 years.

The next chapter describes the exploratory assessment of AI and robotics capabilities using expert evaluations on the selected 13 occupational tasks. It explores ways to address the challenges related to the use of occupational examinations for rating to develop a robust methodological approach for assessing AI and robotics performance in occupations.

References

AkA (2021), *Abschlussprüfung Sommer 2021 Kaufmann/-frau für Büromanagement*, AkA Aufgabenstelle für kaufmännische Abschluss- und Zwischenprüfungen, IHK Nürnberg für Mittelfranken. [25]

Ali, R. et al. (2023), *Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations*, Cold Spring Harbor Laboratory, https://doi.org/10.1101/2023.03.25.23287743. [18]

Antaki, F. et al. (2023), "Evaluating the Performance of ChatGPT in Ophthalmology", *Ophthalmology Science*, Vol. 3/4, p. 100324, https://doi.org/10.1016/j.xops.2023.100324. [19]

Badura, J. (2015), *Handlungsorientierte Aufgaben für schriftliche Prüfungen in der kaufmännischen Berufsausbildung - Erstellung und Korrektur. Leitfaden für Aufgabenersteller/-innen und Korretor/-innen*, [Action-oriented tasks for written examinations in commercial vocational training. A guide for task developers and graders.], AkA Aufgabenstelle für kaufmännische Abschluss- und, Nürnberg. [9]

BIBB (2022), *Verzeichnis der anerkannten Ausbildungsberufe 2022*, (Directory of recognized training occupations 2022), Bundesinstituts für Berufsbildung, https://www.bibb.de/dienst/publikationen/de/17944 (accessed on 27 October 2023). [6]

BIBB (2013), *Empfehlung des Hauptausschusses des Bundesinstituts für Berufsbildung (BIBB) zur Struktur und Gestaltung von Ausbildungsordnungen (HA 158)*, [Recommendations of the Committee of the Federal Institute for Vocational Education and Training for the Structure and Design and Training Regulation Frameworks], Bundesanzeiger Amtlicher Teil (BAnz AT 13.01.2014 S1), http://www.bibb.de/de/32327.htm (accessed on 27 October 2023). [13]

Buckendahl, C. (2017), "Credentialing", in *Testing in the Professions*, Routledge, New York, https://doi.org/10.4324/9781315751672-1. [12]

Cedefop (2020), *Vocational education and training in Germany: Short description*, Publications Office of the European Union, Luxembourg, http://data.europa.eu/doi/10.2801/329932 (accessed on 27 October 2023). [7]

Eulerich, M. et al. (2023), "Can Artificial Intelligence Pass Accounting Certification Exams? ChatGPT: CPA, CMA, CIA, and EA?", *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.4452175. [22]

European Commission (2008), *NACE Rev. 2 – Statistical Classification of Economic Activites in the European Community*, Office for Official Publications of the European Communities, Luxembourg, https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF (accessed on 27 October 2023). [16]

Hambrick, D. and R. Hoffman (2016), "Expertise: A second look", *IEEE Intelligent Systems*, Vol. 31/4, pp. 50-55, https://doi.org/10.1109/mis.2016.69. [14]

ILO (2012), *International Standard Classification of Occupations: ISCO-08*, International Labour Organization, Geneva. [15]

Johnston, J. et al. (2014), "Determining BACB examination content and standards", *Behavior Analysis in Practice*, Vol. 7/1, pp. 3-9, https://doi.org/10.1007/s40617-014-0003-6. [1]

Katz, D. et al. (2023), "GPT-4 Passes the Bar Exam", *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.4389233. [23]

Kleiner, M. and E. Vorotnikov (2017), "Analyzing occupational licensing among the states", *Journal of Regulatory Economics*, Vol. 52/2, pp. 132-158, https://doi.org/10.1007/s11149-017-9333-y. [5]

Koumenta, M. and M. Pagliero (2017), "Measuring prevalence and labour market impacts of occupational regulation in the EU", report for European Commission, Directorate-General for Internal Market, Industry, https://ec.europa.eu/docsroom/documents/20362/attachments/1/translations/en/renditions/native (accessed on 27 October 2023). [4]

Lin, J. et al. (2023), "Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination", *Eye*, https://doi.org/10.1038/s41433-023-02564-2. [20]

National Center for O*NET Development (n.d.), *O*NET 27.2 Database*, https://www.onetcenter.org/database.html (accessed on 24 February 2023). [28]

NCSL (2022), *The National Occupational Licensing Database*, (database), https://www.ncsl.org/labor-and-employment/the-national-occupational-licensing-database (accessed on 30 August 2023). [10]

Noever, D. and M. Ciolino (2023), "Professional Certification Benchmark Dataset: The First 500 Jobs For Large Language Models". [17]

Nori, H. et al. (2023), "Capabilities of GPT-4 on Medical Challenge Problems". [21]

OECD (2023), *Is Education Losing the Race with Technology?: AI's Progress in Maths and Reading*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/73105f99-en. [3]

OECD (2021), *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/5ee71f34-en. [8]

PAL (2019), *Abschlussprüfung Teil 1 - Technische(r) Produktdesigner(in) - Produktgestaltung und -konstruktion, Prüfungsprodukt, Musterprüfung*, PAL Prüfungsaufgaben- und Lehrmittelentwicklungsstelle, IHK Region Stuttgart. [27]

PAL (2014), *Fachkraft für Metalltechnik - Leitfaden für die Zwischenprüfung - Musterprüfung*, PAL Prüfungsaufgaben- und Lehrmittelentwicklungsstelle, IHK Region Stuttgart. [24]

Rüschoff, B. (2019), *Methods of competence assessment in vocational education and training (VET) in Germany – A systematic review*, Bundesinstitut für Berufsbildung, https://www.bibb.de/dienst/publikationen/de/17861 (accessed on 27 October 2023). [2]

US Department of Labor (2021), *Careeronestop*, website, https://www.careeronestop.org/ (accessed on 30 August 2023). [11]

Zahnärztekammer Niedersachsen (2021), *Infos für Auszubildende und Ausbilder - Musterprüfungen*, https://zkn.de/praxis-team/zan-beruf-und-bildung/ausbildung-zfa/infosauszubildende-ausbilder.html (accessed on 10 November 2021). [26]

Notes

[1] https://www.ihk-aka.de/

[2] Prüfungsaufgaben- und Lehrmittelentwicklungsstelle (PAL) https://www.ihk.de/stuttgart/pal

[3] Zentralstelle für Prüfungsaufgaben https://www.ihk-zpa.de/

[4] https://zfamedien.de/zfa/