10. An occupational taxonomic approach to assessing AI capabilities

David Dorsey, Human Resources Research Organization (HumRRO)

Scott Oppler, Human Resources Research Organization (HumRRO)

This chapter proposes an approach for comparing capabilities of human and artificial intelligence (AI) based on comprehensive occupational taxonomies. After a summary of general methodological recommendations, it describes four major steps of the proposed approach. As the first step, it discusses the identification of an occupational taxonomy and its requirements. Second, it proposes a strategy for sampling occupations from the taxonomy. Third, it provides guidance on collecting expert judgement on AI capabilities with regard to the selected occupations. Fourth, it considers the implications of data analysis from expert interviews.

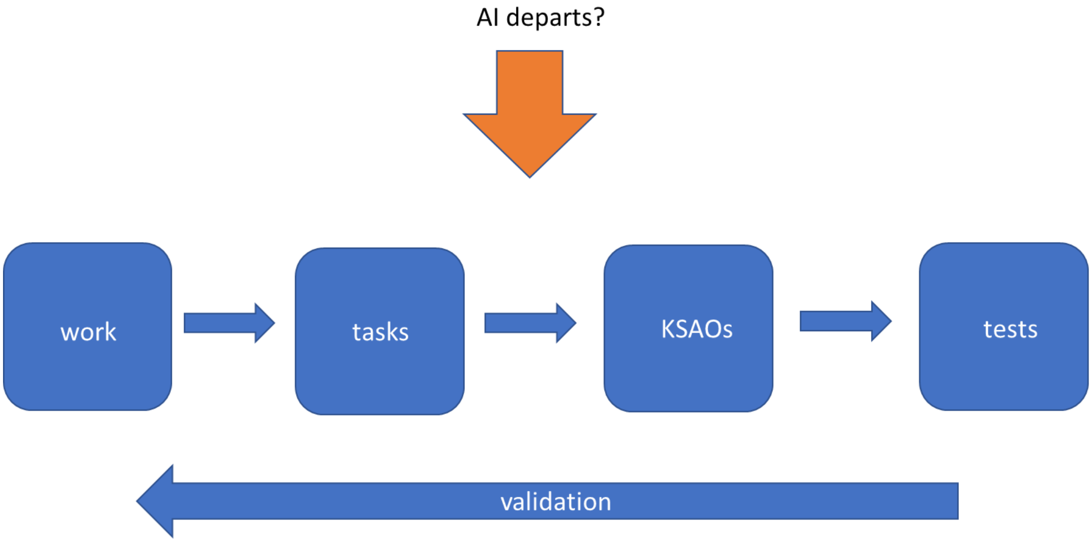
Introduction

This chapter proposes a four-step approach for comparing capabilities of human and artificial intelligence (AI) based on comprehensive occupational taxonomies. To that end, it looks to identify an appropriate taxonomy of human capabilities to assess and to identify available human assessments. After considering assumptions about design, it proposes a four-step approach to evaluate the capabilities of AI and robotics with regard to the world of work.

"Work," "skills," "tasks," etc. are often used in fundamentally diverse ways. Assessment developers typically start with a specific definition of work, usually via a job or practice analysis. This allows understanding of the essential underlying tasks (specific work behaviours) that comprise an occupation, job or position (Figure 10.1). Based on the tasks, one infers the underlying knowledge, skills, abilities and other key characteristics (KSAOs, or worker characteristics) needed to perform these tasks. These KSAOs then serve as the direct targets for assessment development.

Assessment results can only be "valid" to the extent that they support inferences or linkages back to performing actual job-relevant behaviours. Therefore, developers use a variety of validation techniques to gather and form data-based validity arguments. In this way, they complete the inference chain back to "work".

Figure 10.1. An understanding of the work-related assessment development inference chain



To build a regular programme of assessments that track development of AI capabilities and compare them to the distribution of human capabilities, over time, two related problems need to be solved: identifying an appropriate taxonomy of human capabilities to assess; and identifying available human assessments. It will also be necessary to sample occupations and work descriptors (rather than assessing all of them) and sample assessments from a limited set of occupations (based on the availability of relevant assessments).

Assumptions and method design considerations

There are several assumptions and key design considerations.

Subject matter experts will be required

Subject matter experts (SMEs) are needed to judge whether AI can successfully perform a given work activity and/or correctly answer a particular assessment item. Sets of SMEs could include a combination of computer scientists, industrial-organisational psychologists and incumbents in the occupations to evaluate the capabilities of AI systems with regard to these elements.

There is an implicit major "levels of analysis" question

A major "levels of analysis" question is implicit in the methodology. Both human capabilities and AI technologies change in meaning and specificity as descriptions move from the abstract to the concrete. In the United States, for example, the Department of Labor's Occupational Information Network (O*NET)[1] is the official system for describing work at the national level. It describes the same type of work activity in general terms, somewhat more specific terms or very job-specific terms (Figure 10.2). Specific constructs implied in the activity can change at various levels of description.

This same level of analysis question also applies to other work- and worker-oriented descriptors such as skills, knowledge and even aspects such as personality. At possibly the most specific level of analysis are individual assessment items (such as those shown in Appendix A). These are used to assess competence for hiring, advancement, credentialing, etc. within occupations.

Within occupations, there are various career stages or career levels, often described as entry, journeyman, full performance and expert. Thus, the requirements of occupations change across career levels. The measurement of career levels and "career paths" is in and of itself a discrete area of practice and science [e.g. (Carter, Cook and Dorsey, 2011[1])].

The capabilities of artificial intelligence can be judged against many different work-related descriptors

In addition to determining the level of work description used for this research, work can be described in terms of a variety of descriptor types, including:

- job characteristics (e.g. tasks, activities)
- worker characteristics (e.g. abilities, skills)
- occupational assessments (e.g. items on credentialing exams).

Within the study of occupations, pre-employment conditions or requirements might correlate with "lower-level skills". For example, the US Social Security Administration uses a set of "activities of daily living" (ADLs) to determine whether someone in a disability status can work. These include common activities that any self-sufficient person may be expected to perform, such as grocery shopping, dressing and going to work. There are various measurement approaches for determining the requirements of work in terms of such ADLs.

Figure 10.2. Levels of analysis in work descriptors

O*NET Generalized Work Activity Example

- Establishing and Maintaining Interpersonal Relationships — Developing constructive and cooperative working relationships with others and maintaining them over time.

O*NET Detailed Work Activity Example

- Liaise between departments or other groups to improve function or communication.

O*NET Job Task Example

- Represent organization at personnel-related hearings and investigations.

Source: www.onetonline.org/.

Significant trade-offs exist between "fidelity" and "generalisability" across occupations

Within this research challenge, there is a trade-off between "fidelity" and "generalisability". Specifically, the more fine-grained the information about a given occupation is, the larger the amount of this information gets, but it becomes harder to generalise to other occupations. For example, items on occupational credentialing exams tend to be specific with respect to occupations. This makes such exams an intuitive and presumably well-grounded basis for judging work-related capabilities pertaining to an occupation.

However, the judgements associated with these exams have limited application to other occupations. In contrast, abilities and skills tend to be less specific to occupations. Therefore, judgements associated with them apply more readily to other occupations, given the required level of abilities and skills in the other occupations is known.

As discussed above, job characteristics can range from specific tasks to more generally defined work activities. The latter allows judgements to apply across a wider range of occupations, assuming the required level in the other occupations is known. Further, the more occupationally specific the elements included in a judgement task are, the more judgements are required for a given occupation.

Conversely, the less occupationally specific the elements included in the judgement task are, the fewer judgements are required for a given occupation. The challenge, then, is to determine the minimum level of occupational specificity needed to produce useful information about AI capabilities for a given occupation, while maximising generalisability.

It is appealing to assume the "right" level of granularity of information for making decisions can be determined in advance. However, collective experience suggests this level depends heavily on the purpose and context of the decisions, and the ultimate use of the information. With respect to testing AI, there is limited prior research, calling for different approaches, via pilot testing, cognitive lab testing, etc. These are discussed below.

Worker-oriented characteristics (and related job performance) can also be divided into “can do” aspects. These are typically based on knowledge, skill or ability, whereas “will do” aspects focus on non-cognitive aspects such as motivation, personality, stress tolerance, etc. (Borman et al., 1991[2]).

To put a fine point on the trade-offs mentioned above, specifically around the number of judgements needed, consider that O*NET currently contains information regarding requirements for nearly 1 000 occupations. Taking as a prototypical credentialing exam, a given form of the widely known Society for Human Resource Management certification exam presents examinees with 160 items, 96 knowledge items and 64 situational judgement items.[2]

Thus, across all occupations, credentialing exams of this type would reach an astronomical number of judgements (approximately 1 000 x 160 = 160 000), if they were used to assess all possible AI capabilities. Each of these approximately 160 000 items are likely to be specific to the occupation in question, affording little generalisability to other occupations.

One alternative would be indirect generalisability via careful sampling of occupations. In this case, the relevance of results to other occupations could be inferred. However, if the goal is to make inferences about an entire economy, claims would be stronger through a direct form of generalisability evidence.

Higher-level work descriptors, such as those on O*NET, could be a direct mechanism for generalising judgements about capabilities across occupations. Such an approach would not negate the use of specific credentialing assessment items. Instead, a methodology with various levels of information and judgements could gauge what can be gleaned from different levels of analysis. Moreover, using this methodology, specific assessment items can be linked to higher level descriptors as is often done in various forms of content validation.

Proposed approach

Given the trade-offs noted above, a four-step approach is proposed to evaluate the capabilities of AI and robotics with regard to the requirements of the world of work.

Step 1: Identify an occupational information system that includes work-related descriptors or elements representing a range of levels of occupational specificity

Identify an occupational information system. This needs to specify the work and worker requirements for a wide range of occupations. It should also specify the descriptors at different levels of occupational specificity, ranging from the specific to the general. O*NET could be a useful “content model” as it is made up of several different taxonomies regarding occupational characteristics and worker requirements.

O*NET has two characteristics that make it particularly relevant. First, it describes occupations in terms of the knowledge, skills and abilities required of workers in those occupations. Second, it describes how the work is performed in terms of both occupationally specific tasks and work activities at three different levels of specificity (Detailed, Intermediate and Generalised Work Activities). O*NET also has links to Europe’s European Skills, Competences, Qualifications and Occupations (ESCO).[3]

Step 2: Identify a sample of occupations representing a range of job families (e.g. manufacturing, health care)

Identify a sample of occupations to include in the SME judgement workshops (described in Step 3). As noted previously, the number of occupationally specific judgements required for a single occupation can be quite large. This is especially true if the rated stimuli are individual test items on an occupational credentialing exam (which are often in the range of 150 items or more). Therefore, the number of occupations that can be feasibly included in the research is limited.

The proposed approach selects occupations as exemplars of broader job families, allowing most of the workforce to be represented, at least to some extent. For example, the near 1 000 occupations in the O*NET database are classified into 23 job families, as well as 16 broader career clusters. Therefore, one occupation could represent each job family (or career cluster) to help cover the range of occupations in today's workforce. It would select occupations with existing (and available) credentialing exams. They would then be included in the expert judgement tasks as part of Step 3.

Even working with only 16 or 23 occupations (depending on whether they have been sampled from career clusters or job families) would require a sizeable number of judgements on the credentialing exam items (i.e. 16 x 160 = 2 560). However, this is certainly more manageable than attempting to collect such judgements for all occupations in the workforce. Of course, if necessary, this number could be further reduced by grouping the job families or career clusters into a smaller number.

Step 3: Convene subject matter experts to judge the capabilities of AI with respect to different sets of descriptors ranging in degree of occupational specificity

The third step concentrates most of the effort. It collects judgements from SMEs regarding the capabilities of AI with respect to the various sets of descriptors in the research. There are two subcomponents to this step: determining which sets of descriptors and other stimuli (e.g. credentialing exam items) to collect judgements about; and collecting the judgements.

Determining descriptors and stimuli

The collection of judgements should represent a range of levels of occupational specificity to evaluate the trade-off between fidelity and generalisability. To that end, inferences supported by descriptors at different levels of occupational specificity should be compared to identify the level(s) that best manage the trade-off (i.e. providing the greatest fidelity, while still enabling generalisability of judgements across the greatest numbers of occupations).

Using O*NET, for example, SMEs could judge AI capabilities with respect to occupational-specific tasks, as well as at the progressively less occupationally-specific work activities (Detailed, Intermediate and Generalised). There are over 20 000 occupationally specific tasks in the O*NET system. However, only those associated with the occupations identified for the sample (as identified in Step 2) would be included in the proposed data collection. In comparison, there are approximately 2 000 Detailed Work Activities, 300 Intermediate Work Activities and 23 Generalised Work Activities, all of which would be included in the proposed research.

In addition to collecting judgements for items on occupational-specific credentialing exams, it would be possible to collect judgements with regard to O*NET's cross-occupational Abilities, Skills and Knowledge descriptors. The job requirement scales for each of these descriptors are defined in terms of tangible work behaviours (see Appendix A).

Altogether, 52 abilities, 35 skills and 33 knowledge areas can be included in the data collection. Finally, the research should consider inclusion of descriptors representing ADLs discussed earlier. Although not included in O*NET, taxonomies and descriptors for ADLs exist in other sources (Edemekong et al., 2019[3]).

Collecting judgements

Multidisciplinary teams of SMEs should collaborate in judgements concerning AI capabilities with respect to the selected descriptors. These teams should comprise computer scientists with expertise in AI and robotics, industrial-organisational psychologists with expertise in job analysis and human performance, and job incumbents employed in occupations identified in Step 2. Each group would bring different, and important, perspectives regarding the judgements being gathered.

Judgement tasks could vary depending on the descriptors being considered by the SMEs. For example, judgements regarding the capabilities of AI systems to respond correctly to items on the occupational-specific credentialing exams might use an approach similar to that used to assess computer capabilities to respond correctly to items on the Survey of Adult Skills (PIAAC) (Elliott, 2017[4]).

Alternative procedures might be used for judging the capabilities of AI with regard to the work-related tasks and activity descriptors or the abilities, skills and knowledge areas in the O*NET Content Model. For instance, SMEs could identify the maximum point on each O*NET scale to represent the level of activities at which an AI system could be expected to perform, given a specified amount of resources required for development.

These judgements would be similar to those collected in standard-setting studies using the Bookmark method (Karantonis and Sireci, 2006[5]). In that case, SMEs are asked to identify the most difficult item on an assessment that a minimally qualified examinee would be likely to answer correctly.

These judgements would not be tied to specific occupations. Therefore, they would not necessarily need to be collected for each individual occupation included in the sample. That said, job incumbents from a variety of occupations should be included in the teams. They would provide the judgements for these cross-occupational descriptors. They could also collect judgements regarding these descriptors from different groups of SMEs (each focusing on different occupations) to evaluate the extent to which they generalise across occupations.

Regarding the occupationally specific tasks associated with occupations in the sample, SMEs could estimate the level of effort required to develop an AI system to replace (and/or potentially assist) human workers in performing a given task, activity or work behaviour.

The various approaches to eliciting these judgements (using some combination of cognitive laboratories and pilot testing) should be evaluated before embarking on any full-scale effort to collect data.

Step 4: Compare results associated with the different sets of descriptors (or combinations thereof)

The final step would be to analyse the data collected from the SMEs to achieve two goals.

Determine occupations primed to be replaced or aided by AI

First, the specific occupations included in the sample would be evaluated. This would examine the extent to which all or portions of each occupation are primed to be replaced and/or aided by AI technology.

Results would not likely suggest that AI could completely replace workers in any of the occupations in the research. However, they might point to particular types of tasks and activities that may no longer require human workers. More likely perhaps, they could identify the skills and abilities needed by human workers, given the assistance that AI might provide.

Comparing results across the admittedly limited sample of occupations may lead to some valuable inferences regarding the prevalence of these potential changes. Identifying specific “job components” that could be done by AI would allow a search for these job components across a database like O*NET. This, in turn, could possibly generalise the results to new occupations. This is similar to how “job components” are used in synthetic approaches to validation [e.g. (Johnson and Carter, 2010[6])].

Evaluate judgements of AI capabilities

Second, and perhaps more importantly, a data analysis would evaluate the extent to which the different sets of judgements make similar conclusions about AI capabilities.

On the one hand, results associated with the more general descriptors could lead to similar conclusions as the more specific descriptions. In other words, descriptors such as skills; detailed and intermediate work activities could have the same conclusions as occupational-specific tasks and items on occupational-specific credentialing exams. This would suggest that more general judgements could be applied to the other occupations.

On the other hand, results may indicate that inferences associated with descriptors representing different levels of occupational specificity are not sufficiently similar to one another. This would suggest the continued need to collect judgements with regard to occupationally specific descriptors (job-specific tasks; items on credentialing exams) for those occupations of interest not included in Step 3.

The psychometric quality of SME judgements is an additional consideration. The four steps do not include a separate validation of the SME judgements regarding the capabilities of AI to perform various tasks and activities.

However, the data collection could be used to estimate the extent to which different SMEs (or different groups of SMEs) provide similar judgements about different sets of descriptors. For example, judgements regarding the more occupationally specific descriptors may demonstrate greater levels of inter-rater agreement than do the more general descriptors.

Alternatively, the level of agreement among the ratings for the more general descriptors may be relatively high among SMEs in the context of a given occupation (or family of occupations). Yet these judgements may differ for the same descriptors collected in the context of different occupations. The design for data collection strategies needs to consider these possibilities.

Consistency of judgements is a necessary aspect of psychometric quality but not enough to establish validity. To truly establish validity, these judgements must be compared with how AI systems carry out the activities that are the subject of the judgement task. Moreover, this should take place in the context of the specific occupations for which the judgements are being made.

This process could potentially be expensive for occupations new to AI systems. However, it could be possible to include several occupations in the sample for which automated systems have begun to proliferate. Judgements about these occupations and systems could then be compared to judgements about occupations without automation.

References

Borman, W. et al. (1991), “Models of supervisor job performance ratings”, *Journal of Applied Psychology*, Vol. 76, pp. 863-872. [2]

Carter, G., K. Cook and D. Dorsey (2011), *Career Paths: Charting Courses to Success for Organizations and their Employees*, Wiley-Blackwell. [1]

Edemekong, P. et al. (2019), “Activities of daily living (ADLs)”, *StatsPearls [Internet]*, https://www.ncbi.nlm.nih.gov/books/NBK470404/. [3]

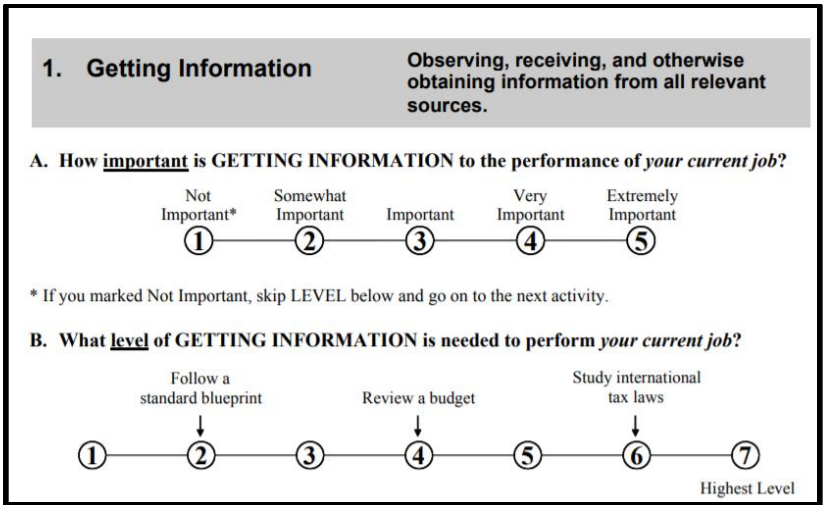
Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264284395-en. [4]

Johnson, J. and G. Carter (2010), “Validating synthetic validation: Comparing traditional and synthetic validity coefficients”, *Personnel Psychology*, Vol. 63/3, pp. 755-795. [6]

Karantonis, A. and S. Sireci (2006), “The Bookmark Standard-Setting Method: A literature review”, *Educational Measurement: Issues and Practice*, Vol. 25, pp. 4-12, https://doi.org/10.1111/j.1745-3992.2006.00047.x. [5]

Annex 10.A. Example rating scales and assessment items

Figure 10.A.1. O*NET GWA rating



1. Getting Information **Observing, receiving, and otherwise obtaining information from all relevant sources.**

A. How <u>important</u> is GETTING INFORMATION to the performance of *your current job*?

Not Important* ① — Somewhat Important ② — Important ③ — Very Important ④ — Extremely Important ⑤

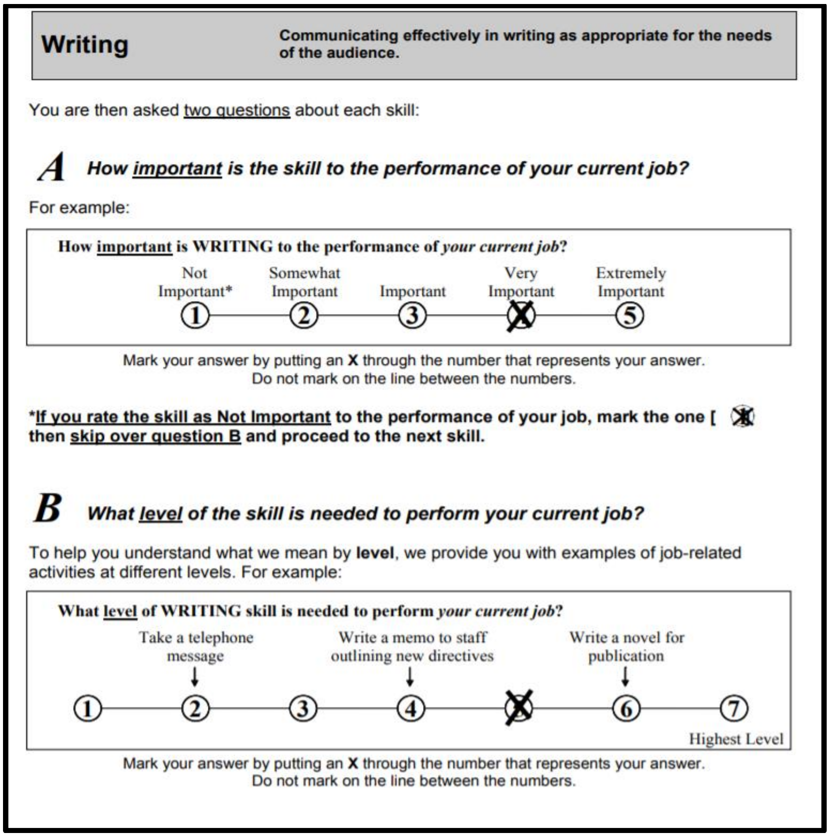
* If you marked Not Important, skip LEVEL below and go on to the next activity.

B. What <u>level</u> of GETTING INFORMATION is needed to perform *your current job*?

① — ② — ③ — ④ — ⑤ — ⑥ — ⑦

Source: http://www.onetonline.org/.

Figure 10.A.2. O*NET skill rating



Writing **Communicating effectively in writing as appropriate for the needs of the audience.**

You are then asked two questions about each skill:

A ***How important is the skill to the performance of your current job?***

For example:

How important is WRITING to the performance of *your current job*?

Not Important*	Somewhat Important	Important	Very Important	Extremely Important
①	②	③	[✓ ④]	⑤

Mark your answer by putting an **X** through the number that represents your answer.
Do not mark on the line between the numbers.

***If you rate the skill as Not Important to the performance of your job, mark the one [☒ then skip over question B and proceed to the next skill.**

B ***What level of the skill is needed to perform your current job?***

To help you understand what we mean by **level**, we provide you with examples of job-related activities at different levels. For example:

What level of WRITING skill is needed to perform *your current job*?

	Take a telephone message ↓		Write a memo to staff outlining new directives ↓		Write a novel for publication ↓	
①	②	③	④	[✓ ⑤]	⑥	⑦ Highest Level

Mark your answer by putting an **X** through the number that represents your answer.
Do not mark on the line between the numbers.

Source: www.onetonline.org/.

Box 10.A.1. SHRM-CP practice question

A small start-up software company realises the technology skill sets of newly hired programmers are more advanced than the existing programmers' skillsets. Recognising the constant business need for these evolving, state-of-the-art skillsets, which is the best workforce development strategy to implement?

- Perform a job redesign for the existing employees that will not require new, updated skills.
- Design a rigorous in-house training programme to get longer-tenured programmers up to speed with the newer programmers.
- Partner with a local community college to offer programmers the opportunity to update their skill sets.
- Offer new hires shorter-term contracts to allow for a continual hiring of programmers with the most up-to-date skills.

Note: SHRM = The Society for Human Resource Management

Box 10.A.2. ASE Mechanic engine repair practice question

Which of the following creates a flapping sound near the front of the engine?

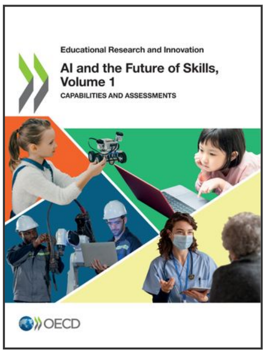
- timing belt tension too tight
- drive belt too tight
- drive belt too loose
- timing belt tension too loose.

Notes

[1] For an overview of O*NET, please see O*NET online at www.onetonline.org/

[2] www.shrm.org/certification/about/descriptions-of-exams/Pages/default.aspx. The full distribution of types of credentialing tests and their tasks in the United States is not known because no comprehensive inventory exists. A recent report from Credential Engine (https://credentialengine.org/counting-credentials-2021/) suggests as many as 967 734 unique credentials in the United States. The report defines “credentials” in a different way than credentialing assessments; yet it does provide a sense of the variety. It highlights “occupational licences” and “occupational certifications”, which numbered about 20 000 in 2020.a

[3] More information regarding the O*NET Content Model is available on the O*NET website at www.dol.gov/agencies/eta/onet.



AI and the Future of Skills, Volume 1

Capabilities and Assessments

Access the complete publication at:
https://doi.org/10.1787/5ee71f34-en

Please cite this chapter as:

Dorsey, David and Scott Oppler (2021), “An occupational taxonomic approach to assessing AI capabilities”, in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/2c1d8961-en