Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

10-Mar-2017

English - Or. English

DIRECTORATE FOR EDUCATION AND SKILLS
CENTRE FOR EDUCATIONAL RESEARCH AND INNOVATION (CERI) GOVERNING BOARD

EDU/CERI/CD/RD(2017)2
Unclassified

DRAFT REPORT ON COMPUTERS AND THE FUTURE OF SKILL DEMAND

4-5 April 2017

This draft report describes the results of an exploratory project to understand current progress in developing these computer capabilities with respect to one set of human skills. The project uses the OECD's Survey of Adult Skills (PIAAC) as a tool for assessing computer capabilities and comparing them with human skills. This exploratory comparison shows how expert knowledge about the capabilities of the technology can be translated into its implications for education.

This room document accompanies the 'Computers and the Future of Skill demands' document [EDU/CERI/CD(2017)4].
This draft version is not to be circulated outside the Governing Board. Comments will be incorporated in a future version that will be released as a publication.

For further information contact:
Mr. Stuart Elliott - SElliott@nas.edu
Mr. Dirk Van Damme, Head of the EDU/IMEP Division - Dirk.Vandamme@oecd.org

JT03410412

English - Or. English

TABLE OF CONTENTS

Tables

Figures

EXECUTIVE SUMMARY

Computer scientists are working on reproducing all human skills and the development of these computer capabilities will have far-reaching implications for work and education. This report describes the results of an exploratory project to understand current progress in developing these computer capabilities with respect to one set of human skills. The project uses the OECD's Survey of Adult Skills (PIAAC) as a tool for assessing computer capabilities and comparing them with human skills. This exploratory comparison shows how expert knowledge about the capabilities of the technology can be translated into its implications for education. Since the exploratory project looks at only one set of work skills, it does not provide a complete basis for forecasting how computers will affect employment and skill demand. However, the project develops an approach that could be used in a comprehensive program to credibly assess computer capabilities across the full set of human work skills and their implications for work and education.

PIAAC measures three general cognitive skills - literacy, numeracy, and problem solving with computers - that are developed during compulsory education and are broadly used by adults both at work and in their personal lives. The questions on the test involve practical problems that would be familiar to most adults who have completed secondary education and live in developed countries. The tasks are designed to be similar to work tasks in many different occupations that require the use of these three skills, even though the test does not assess actual tasks used in any specific occupations.

Extensive research on skills in economics looks at changes in work skills using indirect measures based on wages, education and occupation. This research often finds a pattern of "skill polarization" in employment over the past few decades, with employment increasing for workers with higher and lower skills, while decreasing for workers with mid-level skills. In contrast, this report shows that direct measures of literacy skill suggest a rather different conclusion: compared with two decades ago, there are now more workers using literacy with low or mid-level skills and fewer workers using literacy with high-level skills or not using literacy at all. This pattern results from all workers being more likely to use the literacy skills they have and from somewhat fewer workers having high-level literacy skills.

To understand potential changes in the demand for these skills in the future, the project worked with a group of experts to assess current computer capabilities using the PIAAC test questions. The goal was to identify what questions could be answered by current computer techniques and then to compare that computer performance with the performance shown by adults with different proficiency levels. The exercise focused particularly on computer techniques that have been demonstrated in the research literature but not broadly applied in the workplace. To carry out this analysis, it was necessary to address a number of methodological questions. Given the exploratory nature of the project, the development of this approach for directly comparing human and computer capabilities was itself an important outcome.

The expert assessments generally placed computer capabilities in the middle of the adult proficiency distribution on PIAAC. In all three skill areas, the preliminary results suggest that current computer techniques could perform roughly like adults at Level 2 and that Level 3 performance is close to being possible. In literacy, for example, performance at Level 2 on the 5-level scale means that current computer techniques can generally handle tasks involving several paragraphs of text about a familiar topic and answering questions that require limited inference and enough language understanding to avoid a misleading section of text. However, these techniques cannot reliably answer questions about more difficult passages of text that require more subtle inference and avoiding prominent sections of text that are misleading if not read carefully.

Although the computer capabilities demonstrated in the latest research literature are certainly limited, they suggest that a level of computer performance is now possible in these three skills that is comparable to that of many workers. On average, 31% of the workforce in OECD countries uses one or more of the PIAAC skills on a daily basis at work and has a proficiency in these skills at or below the level of current computer capabilities. Another 31% of the workforce uses these skills on a daily basis and has a proficiency that is close to being possible for computers. Only 13% of the workforce in OECD countries uses the three PIAAC skills on a daily basis and has a proficiency that is clearly beyond the capabilities that computers are close to reproducing.

This exploratory analysis by itself cannot show how computers will affect employment because it looks at only a limited set of work skills. Different mixes of skills are needed for different work tasks and without knowing whether computers have the full set of required skills for a particular task, it is impossible to know whether they would be able to carry it out. For some work tasks, the PIAAC skills will be of primary importance, but for other tasks they will be peripheral or perhaps required in combination with other skills, such as common sense reasoning, expert reasoning, vision, physical movement, or social cognition. A comprehensive program to understand how computers will affect employment must assess these other skills as well.

The exploratory analysis also does not address issues related to the application of these computer capabilities. Studies of technology adoption find that widespread application of a new production technology often takes one or more decades and sometimes never occurs. It takes substantial time and investment for companies to learn about a new technology, decide to adopt it, adapt it for their circumstances, and then implement it across their operations. The current economy reflects the economic impact of computer techniques developed several decades ago, but not the capabilities demonstrated by recent research. The full implications of current computer capabilities for reproducing the PIAAC skills will probably not be seen in the economy for several decades.

Despite the limits of the exploratory analysis, it can provide some preliminary conclusions about the implications for education. During the coming decades, it is likely that there will be strong economic pressure to apply the computer capabilities for the PIAAC skills across the economy. This is likely to reverse the pattern of the recent past of increasing workers using literacy with low and mid-level skills. Without knowing where new applications will be successful, it is reasonable to conclude that there will be an overall decrease in demand for those workers - the vast majority - whose proficiency in the PIAAC skills is no better than that of current computer capabilities. That does not mean that these workers will become unemployed, but they will become less valuable for many work tasks, which will reduce employment in some cases and reduce wages in others.

A standard policy response might advise increased levels of education so that workers are able to move into new types work. However, the exploratory analysis suggests this response may not be viable - at least with respect to the PIAAC skill - because there are no examples of education systems that can prepare the vast majority of adults to perform better on PIAAC than the level computers are close to reproducing. Although some education systems do better than others, those differences are not large enough to help most of the population overtake computers with respect to the PIAAC skills. It is likely that the employment prospects for most adults one or two decades from now will depend increasingly on other types of skills that are not measured by PIAAC. To figure out what policy responses will be helpful in the years ahead, we need to assess computer capabilities across all work skills so we know which skills will be most important for workers to have and can devise appropriate policies to develop those skills.

1 – THE CHALLENGE COMPUTERS POSE TO WORK AND EDUCATION

Computer scientists are working on reproducing all human skills and the development of these computer capabilities will have far-reaching implications for work and education.[1] In response, the structure of the economy and the skills of the workforce will need to be radically transformed over the twenty-first century. Although we cannot know exactly how this transformation will proceed, we can make significant progress in understanding its shape over the next few decades by assessing the full extent of current computer capabilities. In many cases these capabilities have not yet been widely applied. By knowing which computer capabilities are now available and how they relate to human skills, we can better understand which work tasks can potentially be automated in the near future. This understanding can provide the basis for constructing realistic scenarios about the ways that jobs and skill demand will be redefined in the next few decades. This will help policymakers understand how the education system needs to be shaped in turn to prepare today's students for those possible futures.

This report describes an exploratory project to understand computer capabilities with respect to one set of human skills in the context of work and education. The project used the OECD's Survey of Adult Skills (PIAAC) as a tool for understanding the implications of growing computer capabilities. PIAAC measures a set of general cognitive skills - literacy, numeracy, and problem solving with computers - that receive extensive development during compulsory education. Countries invest in developing these skills because they are widely used by adults both at work and in their personal lives: they are "necessary for fully integrating and participating in the labour market, education and training, and social and civic life" and "relevant to many social contexts and work situations" (OECD 2016, p.16). PIAAC measures these skills precisely because of their acknowledged importance both as outputs of the education system and as inputs in the workplace. This report looks at changes in the use of these skills over the past several decades and explores the implications of computers for further changes in the future. The exploratory project described in this report provides the first step towards constructing an ongoing and comprehensive program to assess the capabilities of computers and their implications for work and education.

This report rests on extensive research in economics, with additional perspectives from education and computer science. This first chapter places the current study in the context of those larger literatures and also describes the structure of the report.

Past employment trends in skill demand

The transformation of work skills has been a key aspect of global economic development over the past several centuries. The outlines of this transformation are well-known, involving the long-term shift of employment out of agriculture into manufacturing initially and then ultimately into services, accompanied by large increases in educational attainment. It is helpful to remember the scale of this transformation: in the United Kingdom, for example, employment in services increased from 41% of the workforce in 1890 to 72% in 1998, while average educational attainment increased from 4.8 years to 13.1 years.[2] Different countries are at different stages in this transformation of education and work skills, but the transformation is occurring worldwide and has continuing implications for government policies related to human capital.

The overall shift in employment and increasing demand for education are related to technological change, with new technology during the twentieth century tending on average to increase the demand for

[1] Throughout this report, the term "computers" is used to refer generally to computers, robots, and other types of information and communications technologies.

[2] Employment data: Maddison, 2003, Table 2-24. Education data: OECD, 2014, Table 5.5.

higher skills and decrease the demand for lower skills. This basic historical relationship between technology and education suggests the metaphor of a "race" between technology and education (Goldin and Katz, 2008; Tinbergen, 1974). The general conclusion that technological change is driving the economy towards ever-increasing demands for education is widely accepted. However, the metaphor does not apply to the nineteenth century, when the most salient effect of technological change was to decrease the demand for skills by replacing skilled craft workers with unskilled workers on assembly lines (Acemoglu, 2002).

In the late twentieth century, the change in skill demand became more complex, with the emergence for several decades of a pattern of "polarization", with increasing employment for workers with higher and lower skills and decreasing employment for workers with mid-level skills. This pattern of job change has been found for the United States, many European countries, and Japan (Autor, Katz and Kearney, 2006; Goos, Manning and Salomons, 2009; Ikenaga and Kambayashi, 2010). However, these trends will not necessarily continue. Already, there are questions about whether the pattern of polarized employment changes has continued in recent years, especially with respect to findings of weakness in the demand for workers with higher skills since 2000 (Autor, 2015; Beaudry, Green and Sand, 2016).

Although there are several possible reasons for the trend towards polarization in the labour market, the strongest explanation is that technology is increasingly being used to perform routine tasks that were previously performed by workers (Goos, Manning and Salomons, 2014; Michaels, Natraj and Van Reenen, 2014). According to this explanation, jobs involving routine cognitive tasks typically occur in the middle of the skill distribution and are susceptible to substitution by technology, whereas jobs involving non-routine tasks that cannot yet be carried out by technology occur either at the low or high end of the skill distribution, depending on whether the non-routine tasks require physical or cognitive skills (Autor, Levy and Murnane, 2003).

In general, the research looking at skill polarization uses indirect measures of skill that are available in economic datasets, with nothing like the skill assessments that are used in education research. Typically, "skill" is inferred indirectly from wages, education or occupation. For example, Autor, Katz and Kearney (2006) use three different measures of skill by occupation: mean level of education by occupation; median level of hourly wages by occupation; and data on tasks taken from brief occupational descriptions. Such measures can be useful as rough indicators of skill in the context of economic analyses, but they are far removed from direct measures of skill.

The possibility of technology automating certain skills often leads to questions about changes in the quantity of employment, not just changes in the distribution of skills among the employed. Historically, there have been periodic waves of concern about unemployment resulting from new technology, stretching back at least two centuries to the early industrial revolution (Mokyr, Vickers and Ziebarth, 2015). Despite these waves of concern, unemployment resulting from technological displacement has always been temporary in the past, with increased productivity leading to decreased prices, which lead in turn to increased demand by consumers for both old and new goods, and then to increased demand by employers for workers (Autor, 2015). This adjustment process may not necessarily take place as smoothly as suggested by economic theory (Autor, Dorn and Hanson, 2016) - leaving room for policy interventions to help - but the overall historical experience is one of large-scale and successful redeployment of the workforce as technology shifts the skills needed in different production processes.

In response to several high-profile studies suggesting the possibility of substantial job displacement resulting from computers (Brynjolfsson and McAfee, 2014; Frey and Osborne, 2013), a number of economists have recently looked directly at the employment implications of these technologies. Several studies have specifically looked at the recent employment effects resulting from applications of computers (Bessen, 2015; Falk and Biagi, 2015; Graetz and Michaels, 2015). In general, these studies have found

reassuring and historically-familiar conclusions about the job effects of this technology, with firms, occupations and industries that use higher levels of computers in production experiencing higher productivity and employment. On the other hand, a recent study looking at the effects of trade and offshoring on jobs in the U.S. found that it was computer use in occupations - not trade or offshoring - that led to increased risk of unemployment (Ebenstein, Harrison and McMillan, 2015).

Projecting future trends in skill demand

Looking forward, several recent economic studies have developed theoretical frameworks to explore the potential effects of computers on employment, wages and productivity in the future under various assumptions. These studies extend earlier theoretical work from a period when the technology was substantially less advanced (e.g., Elliott, 1998; Simon, 1977; Zeira, 1998). Acemoglu and Restrepo (2016) develop a model where automation can displace workers in performing older tasks, but where there is endless creation of new and more complex tasks that workers can perform better than machines. Benzell, Kotlikoff, Lagarda and Sachs (2015) develop a model where computers can automate analytical tasks, but not empathetic or interpersonal tasks. Sachs, Benzell, and Lagarda (2015) compare a two-task model, where computers can automate only one of the tasks, with a one-task model where the only task in the economy can be performed by both people and machines. These different models project a variety of results for workers - some good, some bad - depending on assumptions about the economy and public policy, as well as about the fundamental relationship between the capabilities of computers and the skills needed to perform tasks in the economy.

Several studies have used information from computer science to understand the relationship between computer capabilities and the skills and tasks in the economy. The most widely-cited study is by Frey and Osborne (2013), which estimates that 47% of US employment is at a high risk of automation over the next several decades. The Frey and Osborne analysis involves four steps. First, a group of computer scientists classified 70 occupations as either automatable or not, based on a set of job descriptions and their knowledge of current computer capabilities.[3] The occupations chosen were those where the group was most confident in making this judgment. Second, the authors identified job tasks that were most likely to be barriers to computerisation - perception and manipulation, creative intelligence, social intelligence - based on the current state of computer science. Third, occupational data on these hard-to-automate tasks was used to develop a model to predict the automatability classification of the 70 occupations from their tasks and then this model was used to predict the automatability of all occupations in the US economy. Finally, jobs above a predicted automatability of 70% were defined as "high risk" and figures of employment by occupation were used to derive the overall estimate of 47%. Although the model was originally developed for the US, it has been applied to a variety of other countries by substituting different figures for occupational employment in the last step (e.g., Frey, Osborne, and Holmes, 2016; Pajarinen, Rouvinen, and Ekeland, 2015).

Work sponsored by the OECD developed an alternative way of extending the Frey and Osborne automatability judgments for the original 70 occupations to the full economy (Arntz, Gregory and Zierahn, 2016). This work produces an estimate of 9% of jobs in OECD countries as highly automatable, a dramatically lower figure from that estimated by Frey and Osborne from the same starting point. Unlike Frey and Osborne, who use job tasks that are hard to automate as the basis for extending the automatability rating from the 70 occupations to the full economy, the analysis by Arntz, Gregory and Zierahn uses a wide range of job tasks, job characteristics, worker skills and worker characteristics. The information on jobs and workers is obtained from the OECD's Survey of Adult Skills (PIAAC) and includes factors that are not directly related to job tasks, such as gender, age, education, proficiency in literacy and numeracy,

[3] Specifically, the assessments analysed in Frey and Osborne (2013) "were based on eyeballing the O*NET tasks and job descriptions of each occupation" (p. 30).

firm size, and income. In addition, the job task information includes not only factors that are similar to the hard-to-automate tasks identified by Frey and Osborne but other tasks that may not be hard to automate, such as filling forms or calculating percentages. The different result obtained by Arntz and colleagues is explained by the authors as resulting from their "task-based approach", which allows the tasks to vary within occupation according to the job task information in PIAAC.[4]

The McKinsey Global Institute recently issued a report analysing work automatability using an approach that focuses on judgments related to 18 capabilities that are mapped to more than 2,000 work activities in the US and other countries (Manyika et al., 2017). The report estimates that 49% of the activities at work could be automated with current technology. The report does not describe how the judgments related to the 18 capabilities were obtained. The capability judgments were mapped to work activities by a process using the key words in the titles of the work activities.

Another approach to predicting job automatability focuses on worker skills rather than job tasks or activities. Although these two approaches should be complementary, a focus on skills may be more meaningful to the education community. Elliott (2017) uses a sample of recent articles from the computer science literature to identify computer capabilities in four rough skill areas - language, reasoning, movement and vision - that can be mapped to a set of worker descriptors in occupational data for the United States (O*NET). The descriptions of computer capabilities from the research literature are compared to the anchoring tasks on the O*NET scales. The resulting analysis suggests that 82% of current US employment is potentially automatable with the types of capabilities discussed in the current computer science literature.

All of these studies making projections about the potential automatability of current jobs using current technology include caveats about the various economic, institutional and social factors that affect the application of technology. There is an extensive literature about the factors that influence the diffusion of innovation (Rogers, 1995), and when diffusion does occur it can often take several decades or more (Comin and Hobijn, 2010; Griliches, 1957; Mansfield, 1961), even though diffusion speed has increased in recent years. None of the studies of potential automatability focus on an exact timeframe, with all of them referring loosely to computer applications that could potentially happen over a period of several decades. Of course, it is quite possible that some potential applications of current technology will not occur over this period, while it is also possible that even more advanced technology will be developed. All of the studies note that projection of automatability of a percentage of jobs or occupations or tasks over several decades does not mean that the people who currently perform those activities will become unemployed - or even that they will change jobs. Instead, the crucial question relates to the way that their activities and required skills will be redefined, whether or not their job or occupation changes over this period.

Separately within computer science, there is also work to assess the capabilities of current computer techniques with standardized tests developed for people. Early versions of this work go back several decades (O'Neil and Baker, 1994). This literature encompasses several different types of tests, including elementary and secondary school tests in science and math (Clark and Etzioni, 2016); verbal IQ tests for young children (Ohlsson, Sloan, Turán, and Urasky, 2015); and university entrance exams (Arai and Matsuzaki, 2014). To work towards the goal of a broader development and assessment of artificial intelligence, computer scientists have proposed formal assessments in social-emotional intelligence

[4] Arntz and colleagues do not provide any analyses to prove that it is the variation within occupation by tasks that explains the substantial difference in their results from those of Frey and Osborne (2013). Another plausible explanation of their result is that it comes from using a very different set of job features that includes many things that are not job tasks at all. Their report does not include any results for models that use job tasks alone as the basis for the extrapolation.

(Jarrold and Yeh, 2016), physical perception and action (Ortiz, 2016), visual interpretation (Zitnick et al., 2016), and common sense reasoning (Davis, 2016).

Plan for the study and report

This study builds on this prior research related to past and future trends related to work skills by using the OECD's Survey of Adult Skills (PIAAC) to look at the use of general cognitive skills in the workplace. Chapter 2 looks backwards, using PIAAC to describe the distribution of proficiency in the workforce, the use of skill, and the ways these have changed over the past two decades. Chapters 3 and 4 turn from workforce skills to computer capabilities, describing the development and results of an assessment of the capabilities of computers using PIAAC. Chapter 3 describes the development of the approach and Chapter 4 describes the results. These chapters provide a way of understanding a technology that could change skill demand in the decades ahead. Chapter 5 discusses the implications of computer capabilities for the future of the skill changes discussed in Chapter 2 and considers the policy implications for education.

REFERENCES

Acemoglu, D., 2002, Technical Change, Inequality and the Labor Market, *Journal of Economic Literature*, *40*, 7-72.

Acemoglu, D., and P. Restrepo, 2016, The Race Between Man and Machine: Implications of Technology for Growth, Factor Shares and Employment, National Bureau of Economic Research, Working Paper No. 22252.

Arai, N.H., and T. Matsuzaki, 2014, The Impact of AI on Education – Can a Robot Get Into the University of Tokyo?, in C.-C. Liu et al., eds., Proceedings of the 22nd International Conference on Computers in Education, Japan: Asia-Pacific Society for Computers in Education.

Arntz, M., T. Gregory, and U. Zierahn, 2016, The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis, OECD Social, Employment and Migration Working Papers, No. 189, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jlz9h56dvq7-en.

Autor, D.H., 2015, Why Are There Still So Many Jobs? The History and Future of Workplace Automation, *Journal of Economic Perspectives*, *29*, 3, 3-30.

Autor, D.H., D. Dorn, and G.H. Hanson, 2016, The China Shock: Learning from Labor Market Adjustment to Large Changes in Trade, National Bureau of Economic Research, Working Paper 21906.

Autor, D.H., L.F. Katz and M.S. Kearney, 2006, The Polarization of the U.S. Labor Market, *American Economic Review*, *96(2)*, 189-194.

Autor, D.H., F. Levy, and R.J. Murnane, 2003, The Skill Content of Recent Technological Change: An Empirical Exploration, *Quarterly Journal of Economics*, *118(4)*, 1279-1333.

Beaudry, P., D.A. Green, and B.M. Sand, 2016, The Great Reversal in the Demand for Skill and Cognitive Tasks, *Journal of Labor Economics*, *34*, 1, S199-S247.

Benzell, S.G., L.J. Kotlikoff, G. LaGarda, and J.D. Sachs, 2015, Robots are Us: Some Economics of Human Replacement, National Bureau of Economic Research, Working Paper No. 20941.

Bessen, J.E., 2015, How Computer Automation Affects Occupations: Technology, Jobs and Skills, Boston University School of Law, Law and Economics Research Paper No. 15-49. Available at http://www.bu.edu/law/faculty/scholarship/workingpapers/2015.html.

Clark, P., and O. Etzioni 2016, My Computer Is an Honor Student – But How Intelligent Is It? Standardized Tests as a Measure of AI, *AI Magazine, 37(1), Spring: 5-12.*

Comin, D., and B. Hobijn, 2010, An Exploration of Technology Diffusion, *American Economic Review*, *100*(5), 2031-59.

Davis, E., 2016, How to Write Science Questions That Are Easy for People and Hard for Computers, *AI Magazine*, *37*(1), Spring, 13-22.

Ebenstein, A., A. Harrison and M. McMillan, 2015, Why are American Workers Getting Poorer? China, Trade and Offshoring, National Bureau of Economic Research, Working Paper 21027.

Elliott, S.W., 2017, Projecting the Impact of Information Technology on Work and Skills in the 2030s, in J. Buchanan, D. Finegold, K. Mayhew and C. Warhurst, eds., *The Oxford Handbook of Skills and Training*, Oxford University Press.

Elliott, S.W., 1998, Computer Technology, Human Labor, and Long-Run Economic Growth, Heinz School Working Paper 98-23, Carnegie Mellon University.

Falk, M., and F. Biagi, 2015, Empirical Studies on the Impacts of ICT Usage in Europe, Joint Research Centre of the European Commission, Institute for Prospective Technological Studies, Digital Economy Working Paper 2015/14, JRC98693.

Frey, C.B., and M.A. Osborne, 2013, The Future of Employment: How Susceptible are Jobs to Computerization?, Oxford Martin School, September.

Frey, C.B., M.A. Osborne, and C. Holmes, 2016, Technology at Work v2.0: The Future is Not What It Used to Be, Citi GPS and Oxford Martin School.

Goldin, C., and L. Katz, 2008, *The Race Between Education and Technology*, Cambridge, MA: Harvard University Press.

Goos, M., A. Manning, and A. Salomons, 2014, Explaining Job Polarization: Routine-Biased Technological Change and Offshoring, *American Economic Review*, *104(8)*, 2509-2526.

Goos, M., A. Manning, and A. Salomons, 2009, Job Polarization in Europe, *American Economic Review: Papers & Proceedings*, *99(2)*, 58-63.

Graetz, G., and G. Michaels, 2015, Robots at Work, Centre for Economic Performance, CEP Discussion Paper No. 1335.

Griliches, Z., 1957, Hybrid Corn: An Exploration in the Economics of Technological Change, *Econometrica, 25(4),* 501-522.

Ikenaga, T., and R. Kambayashi, 2010, Long-term Trends in the Polarization of the Japanese Labour Market: The Increase of Non-routine Task Input and Its Valuation in the Labour Market, Working Paper, Institute of Economic Research, Hitotsubashi University.

Jarrold, W., and P.Z. Yeh, 2016, The Social-Emotional Turing Challenge, *AI Magazine*, *37*(1), Spring, 31-38.

Maddison, A., 2001, *The World Economy: A Millennial Perspective*, OECD.

Manyika, J., M. Chui, M. Miremadi, M. J. Bughin, K. George, P. Willmott, and M. Dewhurst, 2017, *A Future That Works: Automation, Employment, and Productivity*, McKinsey Global Institute. Available at: http://www.mckinsey.com/global-themes/digital-disruption/harnessing-automation-for-a-future-that-works [accessed 23 January 2017].

Mansfield, E., 1961, Technical change and the rate of imitation, *Econometrica*, *29*(4), 741-66.

Michaels, G., A. Natraj, and J. Van Reenen, 2014, Has ICT Polarized Skill Demand? Evidence from Eleven Countries Over Twenty-Five Years, *The Review of Economics and Statistics*, *96(1)*, 60-77.

Mokyr, J., C. Vickers, and N.L. Ziebarth, 2015, The History of Technological Anxiety and the Future of Economic Growth: Is This Time Different?, *Journal of Economic Perspectives*, *29*, 1, 31-50.

OECD, 2016, *The Survey of Adult Skills: Reader's Companion*, *Second Edition*, OECD Skills Studies, OECD Publishing, Paris.

OECD, 2014, *How Was Life?: Global Well-being Since 1820*, Van Zanden, J.L., et al., eds.

Ohlsson, S., R.H. Sloan, G. Turán, and A. Urasky, 2015, Measuring an Artificial Intelligence System's Performance on a Verbal IQ Test for Young Children, University of Illinois, Chicago.

O'Neil, H.F., and E.L. Baker, eds., *Technology Assessment in Software Applications*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Ortiz, C.L., 2016, Why We Need a Physically Embodied Turing Test and What It Might Look Like, *AI Magazine*, *37*(1), Spring, 55-62.

Pajarinen, M., P. Rouvinen, and A. Ekeland, 2015, Computerization and the Future of Jobs in Norway, Discussion Paper, Research Institute of the Finnish Economy and Statistics, Norway.

Rogers, E., 1995, *Diffusion of Innovations*, 4th edition, New York: Free Press.

Sachs, J.D., S.G. Benzell, and G. LaGarda, 2015, Robots: Curse or Blessing? A Basic Framework, National Bureau of Economic Research, Working Paper No. 21091.

Simon, H.A., 1977, *The New Science of Management Decision*, revised edition, Englewood Cliffs: Prentice-Hall.

Tinbergen, J., 1974, "Substitution of graduate by other labour,"Kyklos 27(2): 217-26.

Zeira, J., 1998, Workers, Machines, and Economic Growth, *Quarterly Journal of Economics*, *113*(4): 1091-1117.

Zitnick, C.L., A. Agrawal, S. Antol, M. Mitchell, D. Batra, and D. Parikh, 2016, Measuring Machine Intelligence through Visual Question Answering, *AI Magazine*, *37*(1), Spring, 63-72.

2 – CHANGES IN SKILLS AND SKILL USE DESCRIBED IN PIAAC

The Survey of Adult Skills (PIAAC) measures a set of general cognitive skills that are developed during education and widely used at work. The survey includes tests of skills in literacy, numeracy and problem solving with computers. The survey also collects information about the ways that adults use skills, as well as various demographic characteristics. This chapter analyses the results of PIAAC to determine what it tells us about changes in skills and skill use over time.

The objective of the chapter is to provide a rough measure of changes in skill demand that can be compared to the economic analyses described in Chapter 1. In general, that literature has analysed shifts in worker skills using data related to worker education, worker pay and occupational activities. In contrast, PIAAC Skills makes it possible to study changes in skill with direct measures of worker skill. Although there are many limitations to these data, they provide a novel perspective on changes in skill demand that can be linked to computer capabilities to consider possible future changes.

After a brief overview of PIAAC, this chapter starts by considering two limitations that need to be taken into account to use the survey to measure change in skill demand over time. It then looks at basic results for skill proficiency and use over the past two decades, producing a picture of changes in skill demand that is somewhat different than the finding of polarization discussed in the economic literature.

Overview of PIAAC

The Survey of Adult Skills, a product of the OECD Programme for the International Assessment of Adult Competencies (PIAAC), assesses the proficiency of adults aged 16-65 in literacy, numeracy and problem solving with computers. In addition to assessing these three competencies, PIAAC also collects information about each respondent's background and context, including information about participation in activities that use the three competencies. The survey is administered by trained interviewers, usually in the respondent's home. It starts with a background questionnaire, which typically takes 30-45 minutes to complete. Each respondent then takes the competency assessment in one or two of the three domains, usually taking about 50 minutes. For further information about the design of the assessment see OECD (2016b).

Two rounds of data collection have been completed so far. In 2011-12, data were collected in 24 countries and economies.

In 21 countries, the entire national population was covered; these countries included Australia, Austria, Canada, Cyprus,[5] the Czech Republic, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Japan, Korea, the Netherlands, Norway, Poland, the Slovak Republic, Spain, Sweden and the United States. In three other countries only part of the population was covered: in Belgium, data were collected in Flanders; in the United Kingdom, data were collected in England and Northern Ireland; in the Russian Federation, data do not cover the Moscow municipal area. In 2014-15, data were collected from an

[5] Note regarding Cyprus: Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue". Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

additional nine countries. In eight countries, the entire national population was covered; these countries included Chile, Greece, Israel, Lithuania, New Zealand, Singapore, Slovenia and Turkey. In Indonesia, data were collected in the Jakarta municipal area only. The total sample includes about 216 000 adults, with national samples ranging from about 4 000 to a maximum of nearly 27 300.

During the process of scoring the assessment, a difficulty score is assigned to each task, based on the proportion of respondents who complete it successfully. These scores are represented on a 500-point scale for each of the three domains. Respondents are placed on the same 500-point scale, using the information about the number and difficulty of the questions they answer correctly. At each point on the scale, an individual with a proficiency score of that particular value has a 67% chance of successfully completing test items located at that point. This individual will also be able to complete more difficult items with a lower probability of success and easier items with a greater chance of success. To help interpret the results, the reporting scales for each domain are divided into a small number of proficiency levels.

Analyses of the PIAAC scores show the close relationship between these skills and labour force outcomes, including wages and employment (OECD, 2016a; Hanushek, Schwerdt, Wiederhold, and Woessmann, 2015). PIAAC scores have also been used to study mismatch between the skills required by a job and the proficiency of workers (e.g., OECD 2016a).

Using PIAAC to measure changes in skill demand

To provide a measure of changes in skill demand over time with PIAAC, there are two challenges that must be solved related to the limitations of the survey data. First, data on worker skill proficiency needs to be transformed into data on the skills that are actually used at work. Second, data from a one-time survey needs to be transformed into longitudinal data. PIAAC provides ways to address both of these challenges.

Data about skill use

One problem with using data on worker proficiency is that the measures indicate skills that are potentially supplied to the workplace but may not actually be used. An extensive literature on skill mismatch indicates that many workers either have skills that may not be used in their job or lack skills that may be important in their job (Quintini, 2011). A direct measure of skill avoids the problem of assuming that workers in a particular job may have a skill that they do not actually possess, but it leaves the problem that their measured skills may not be used. However, PIAAC also provides information on skill use at work. This chapter combines the measures of skill proficiency and skill use to provide a rough measure of skills that are used.

Data about change over time

Although PIAAC is so far only a one-time survey, it was designed to be comparable in some respects to two prior surveys of adult literacy: the International Adult Literacy Survey (IALS) carried out in 1994-1998 and the Adult Literacy and Life Skills Survey (ALL) carried out in 2003-2007 (OECD, 2016b). In order to look at changes over time, the analysis in this chapter combines the results for PIAAC for countries surveyed during 2011-2012 and 2014-2015 with the results for IALS. There are 19 countries or economies that participated in PIAAC that also participated in IALS, with results 13-18 years apart, depending on the country.[6]

[6] The countries or economies participating in both surveys include Australia, Canada, Chile, the Czech Republic, Denmark, England (UK), Finland, Flanders (Belgium), Germany, Ireland, Italy, the Netherlands, New Zealand, Northern Ireland (UK), Norway, Poland, Slovenia, Sweden and the United States. ALL is

Because of changes between the different surveys, there are limits in the ability to compare the results of IALS and PIAAC. The literacy domain in PIAAC incorporates material that was assessed in two separate domains of prose and document literacy in IALS (OECD, 2016b). However, the analysis of the literacy data for PIAAC included a re-analysis of the data from IALS to create scores for a comparable joint literacy domain for the earlier survey (OECD, 2013). Over half of the literacy items used in PIAAC had also been used in IALS, and these linking items provided the basis for constructing comparable scales for the two surveys. It is not possible to compare the other two skill areas assessed by PIAAC. Because the numeracy domain is substantially different than the quantitative literacy domain included in IALS, it is not possible to construct a comparable scale for the earlier survey. For the third domain of problem solving with computers there was no analogous assessment in IALS.

PIAAC and IALS also both ask questions about the use of skills in their respective background questionnaires. There was substantial change in the specific questions and the structure of the possible responses between the two surveys. However, there is sufficient overlap in the design of the two questionnaires that it is possible to identify a small set of questions and response categories that can be compared.

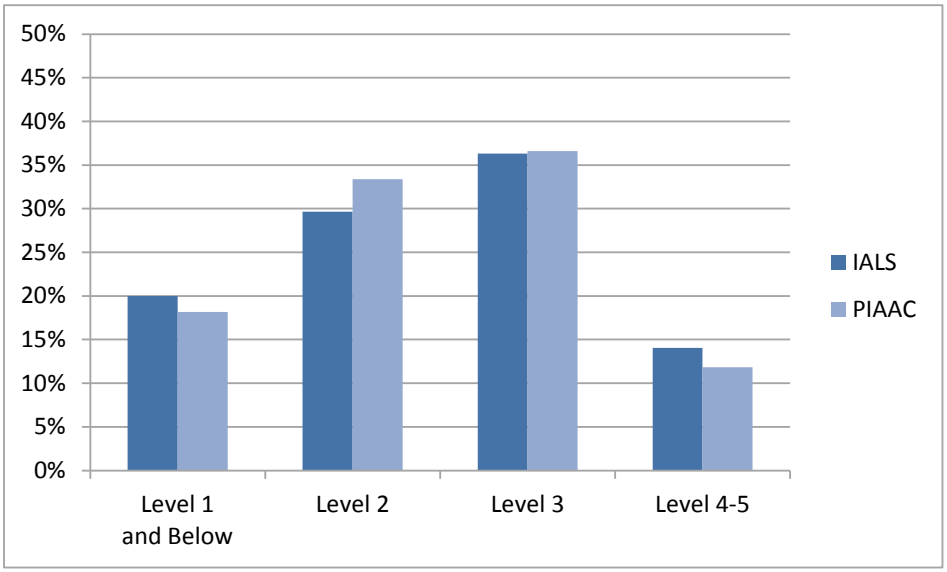
Basic results on literacy proficiency

The results for the adult skill surveys are reported on 500-point scales, which are used to describe both the difficulty of individual test questions and the proficiency of individual adults who took the survey. For ease in understanding, the continuous scales are often described using five proficiency levels. At Level 1 and below, the literacy questions use short texts of a few sentences with basic vocabulary and ask about information that can be clearly identified in the text from the words used in the question. At the higher levels, the texts are longer and the questions may require interpreting or synthesizing the text, as well as avoiding misleading information in the text that may superficially appear to provide the answer. Although the questions at the higher levels are more difficult, the topics are still limited to subjects that are familiar to most adults in developed countries and the material is not technical. For more information on the construction of the literacy test and the content of the questions see OECD (2016b).

Figure 2.1 shows the literacy proficiency results by level, averaged across 19 OECD countries and economies included in both IALS and PIAAC. Because of relatively small numbers of adults at the top and bottom of the scale, those who are Level 1 and below are combined in a single category, as are those at Levels 4 and 5. For both surveys, over two-thirds of the adults are in Levels 2 and 3. In the 13-18 years between the two surveys, the primary change was to increase the proportion of adults at Level 2 by 4 percentage points and to decrease the proportion in the bottom and the top categories by 2 percentage points each.

not used because only 7 of the 19 countries participated in ALL and the shorter time interval provides less opportunity to observe change.

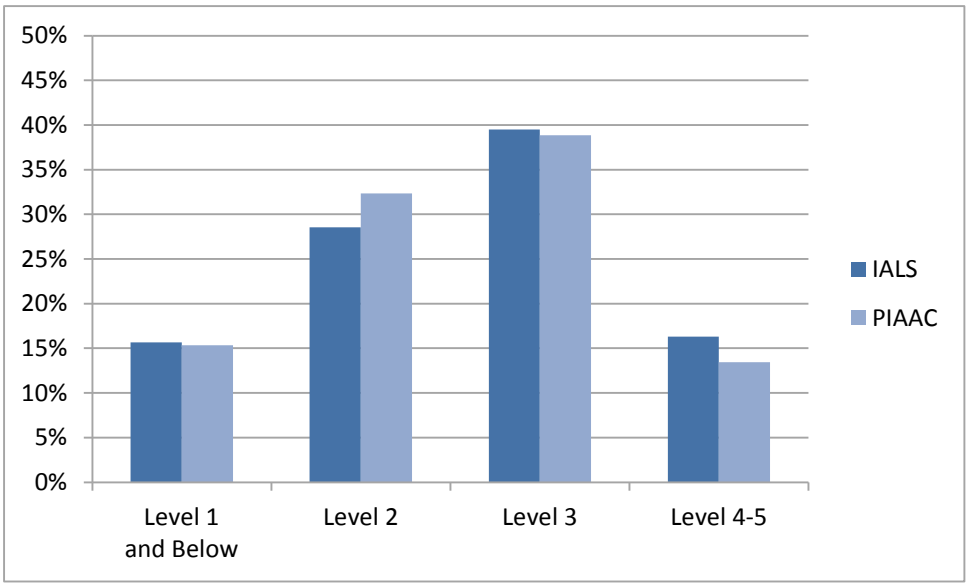
Figure 2.1 Distribution of adult population by level of literacy, IALS and PIAAC



Source: International Adult Literacy Survey (1994-1998), and Survey of Adult Skills (PIAAC) (2012, 2015), Table A2.1.

Of course, many people in the full population are not in the labour force. For comparison, Figure 2.2 shows the literacy proficiency figures for the workforce only. Compared to the full population, the literacy of the workforce is shifted towards higher proficiency levels, with fewer people at Level 2 and below more people at Levels 3-5. However, the change between the two surveys is similar, showing the same increase in the proportion of adults at Level 2 and a slightly larger decrease in the proportion of adults at Levels 4 and 5.

Figure 2.2 Distribution of workers by level of literacy, IALS and PIAAC



Source: International Adult Literacy Survey (1994-1998), and Survey of Adult Skills (PIAAC) (2012, 2015), Table A2.2.

The change in the literacy proficiency between IALS and PIAAC is not well understood. It would be reasonable to expect that the change would be related to the changes in the composition of the population over this period. Countries have generally increased their levels of education over the past several decades, and this would be expected to lead to increasing levels of literacy proficiency over time (OECD, 2016a). At the same time, countries have also experienced increased aging and immigration, both of which are generally associated with lower levels of skill. However, attempts to use these different trends to explain the change in literacy proficiency between IALS and PIAAC with shifts in the composition of the population have not been successful (Paccagnella, 2016). It is also not the case that the trend is specific to a few countries. The increase in the proportion of workers at Level 2 is broadly consistent across countries, with only Chile showing a statistically significant decrease instead of the increase shown on average across countries (Appendix Table A2.2). The decrease in the proportion of workers at Levels 4-5 is somewhat less consistent, but still only three countries (Australia, Poland and Slovenia) show a statistically significant increase in contrast to the decrease shown on average across countries.

Basic results on skill use

Both surveys include a number of questions related to the use of skills at work. In each case, several of these questions concern the use of written material. The wording is quite similar between the two surveys for five questions related to the use of directions, letters, articles, manuals, and diagrams (OECD, 2013).[7] The response categories for these questions are not exactly the same for the two surveys, but the responses in both cases can be aggregated to identify either daily or weekly skill use.[8]

This information on skill use frequency makes it possible to identify the proportion of workers who use their literacy skills as a regular part of their job. Of course, this frequency information alone does not indicate whether the amount of time using literacy is large or small, since even daily use could be for the entire day or only a few minutes. And it does not indicate how important the literacy activity is to the job being performed, since it might be central to the task or involve only some secondary activity, like time-keeping. And it does not indicate how difficult the literacy task is. However, the frequency information does provide an indicator of the proportion of workers whose jobs involve some regular use of literacy skills.

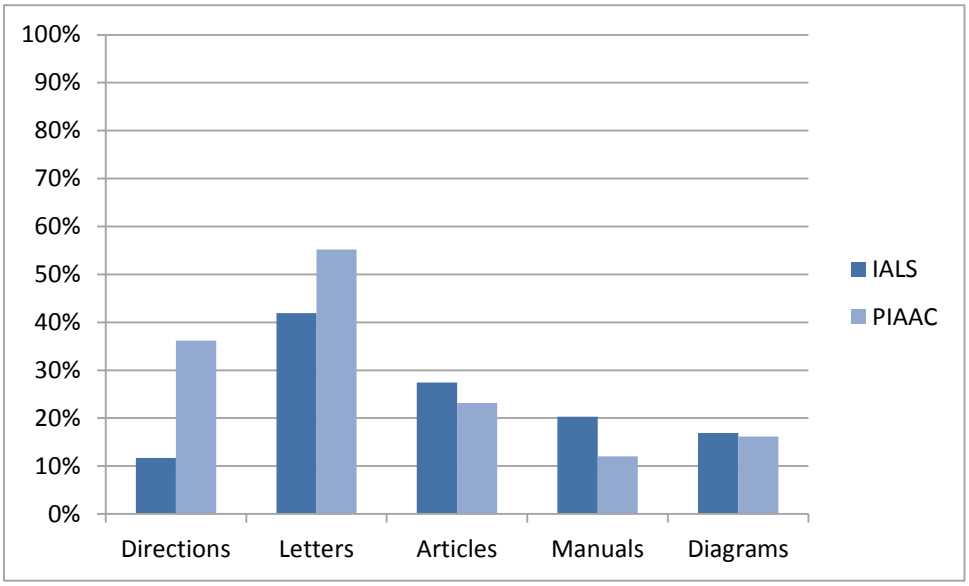
Figure 2.3 shows the portion of the workforce using each of the skills on a daily basis, averaged across 18 OECD countries and economies included in both IALS and PIAAC.[9] The figure shows that the daily use of both directions and letters increased substantially during the 13-18 years between the two surveys, while the other three types of materials show modest decreases. Figure 2.4 shows the change with respect to weekly use for the same set of skills, with a similar pattern except that the increase in the use of letters is smaller and the other changes are larger.

[7] For directions, IALS asks about "directions or instructions for medicines, recipes, or other products," whereas PIAAC asks about "directions or instructions." For letters, IALS asks about "letters or memos," whereas PIAAC asks about "letters, memos or e-mails." For articles, IALS asks about "reports, articles, magazines or journals," whereas PIAAC asks about "articles in newspapers, magazines or newsletters." For manuals, IALS asks about "manuals or reference books, including catalogues," whereas PIAAC asks about "manuals or reference materials." For diagrams, IALS asks about "diagrams or schematics," whereas PIAAC asks about "diagrams, maps or schematics."

[8] For IALS, the response categories are "every day, a few times a week, once a week, less than once a week, rarely or never" (OECD, 2013). For PIAAC, the response categories are "never, less than once a month, less than once a week but at least once a month, at least once a week but not every day, every day."

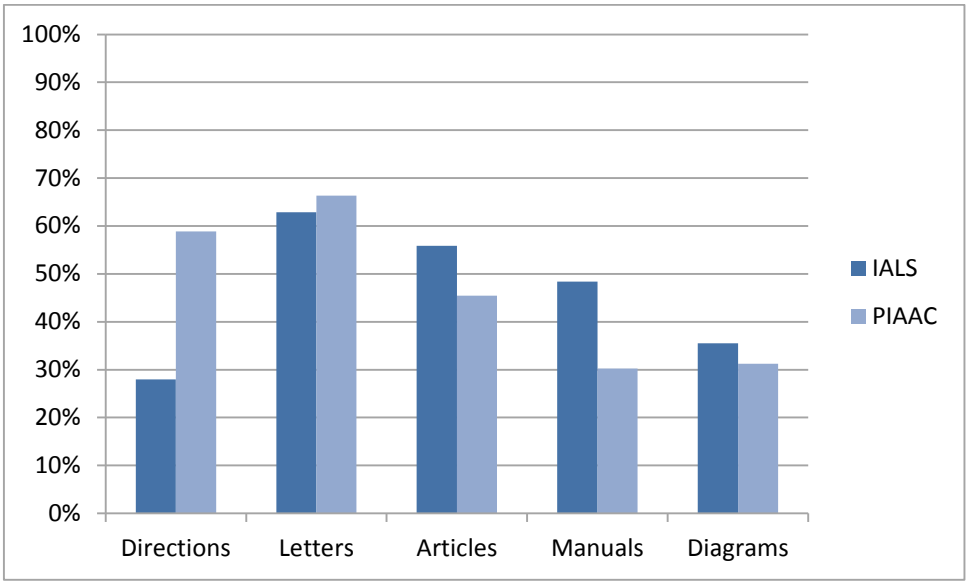
[9] IALS data on daily skill use are not available for Australia.

Figure 2.3 Daily use of different written materials at work, IALS and PIAAC



Source: International Adult Literacy Survey (1994-1998), and Survey of Adult Skills (PIAAC) (2012, 2015), Table A2.3.

Figure 2.4 Weekly use of different written materials at work, IALS and PIAAC

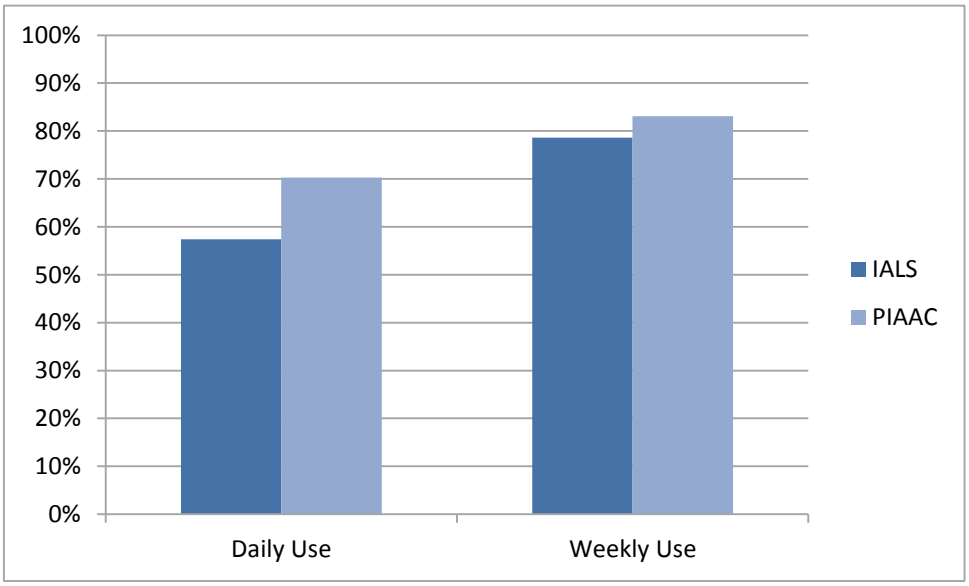


Source: International Adult Literacy Survey (1994-1998), and Survey of Adult Skills (PIAAC) (2012, 2015), Table A2.4.

There is no information available to indicate how adults interpreted the descriptions of these different types of written material. Some of the changes in indicated use between the two surveys may be due to the small differences in wording. However, considering common sense meanings for the different types of materials indicated, the patterns in Figures 2.2 and 2.3 suggest some change towards shorter and less complex written materials between the two surveys: directions are often shorter than manuals, and diagrams are often somewhat complex.

For the purposes of knowing whether literacy skills are being used at all, and without having more detail about how adults understood the different descriptions of written material, it is useful to aggregate the results across the different types of written materials. Relatively comparable literacy proficiency could be used on each of the different types. To aggregate across the different literacy skill use questions, Figure 2.5 shows the proportion of workers who use at least one of the five types of written materials at work on a daily or weekly basis. The figure shows a substantial increase of 13 percentage points in the proportion of workers using written materials on a daily basis, along with a more modest increase of 4 percentage points for use on a weekly basis. Although there are differences across countries, all countries show a statistically significant increase between the two surveys in the proportion of workers using these written materials on a daily basis, with the increase ranging from 4 percentage points for Italy to 26 percentage points for Ireland (Appendix Table A2.5). The changes in weekly use are more mixed across countries, with three countries (Denmark, Germany and Italy) showing decreases in weekly use, ranging from 3 to 9 percentage points. All other countries show increases, ranging from 1 percentage point for Finland to 19 percentage points for Poland.[10]

Figure 2.5 Daily and weekly use of any written materials at work, IALS and PIAAC



Source: International Adult Literacy Survey (1994-1998), and Survey of Adult Skills (PIAAC) (2012, 2015), Table A2.5.

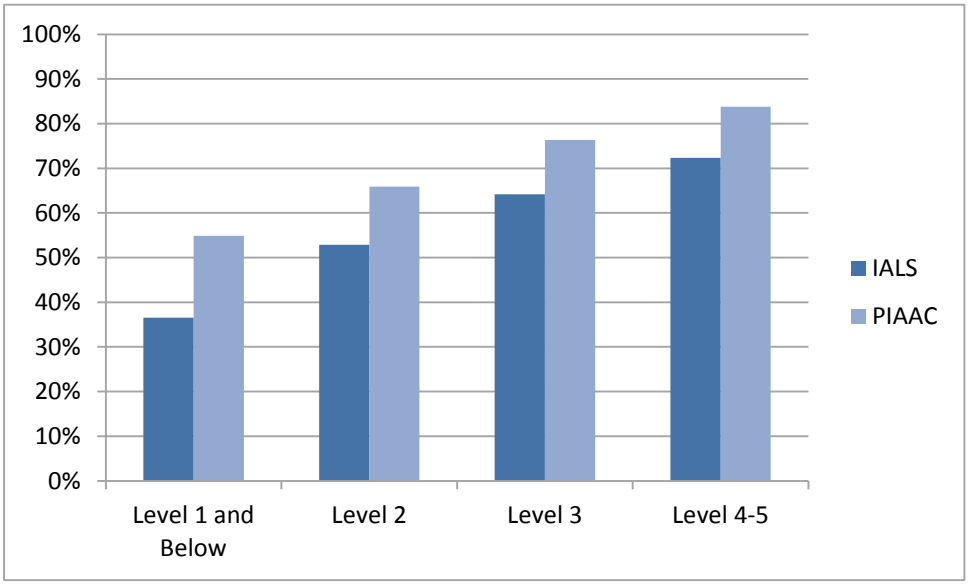
Utilized literacy proficiency – combining results on literacy proficiency and use

Overall, the rates of use are higher for workers with higher skill levels and those rates have increased between the two surveys. Figure 2.6 shows the proportion of workers at each literacy proficiency level who use their skills on a daily basis, with the increases ranging from 18 percentage points for workers with proficiency at Level 1 and Below, to 11 percentage points for workers at Levels 4 and 5. Figure 2.7 shows the same proportions for workers who use their skills on a weekly basis. Although all proficiency levels again show increases in the rates of use, those increases are small except for the lowest proficiency levels.

[10] Quintini (2016) conducts a similar analysis with respect to weekly skill use with the IALS and PIAAC data. She finds similar results as Figure 2.4 when use is considered separately by type of material. To aggregate across types of material, she averages the frequency ratings across the different measures, which produces a finding of no change over time, in contrast to the increase found here by using the maximum frequency across the different measures. She does not analyse daily skill use.

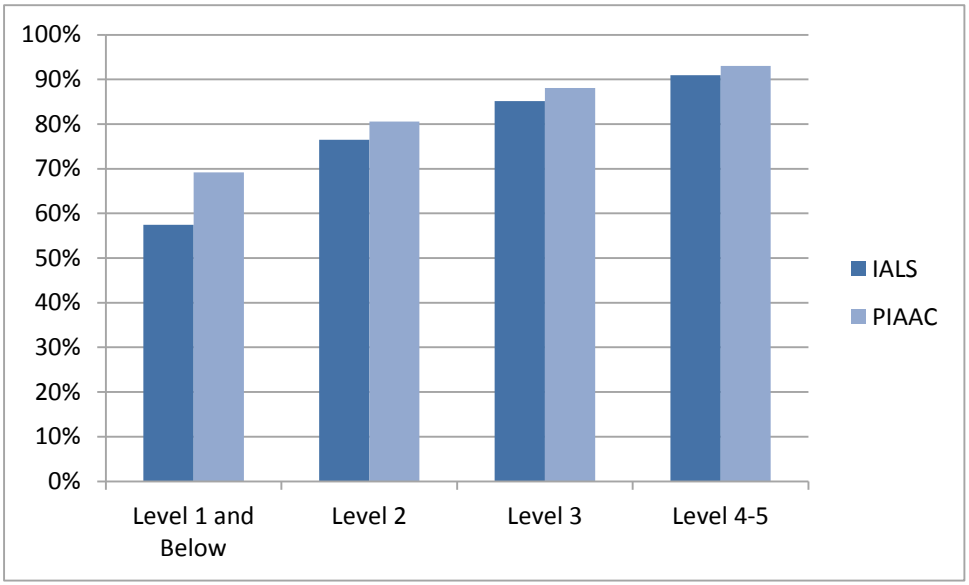
For the workers at the higher proficiency levels, there is little room for further increases because most of these workers use their literacy skills on a regular basis.

Figure 2.6 Proportion of workers at each proficiency level who use literacy skills daily, IALS and PIAAC



Source: International Adult Literacy Survey (1994-1998), and Survey of Adult Skills (PIAAC) (2012, 2015), Table A2.6.

Figure 2.7 Proportion of workers at each proficiency level who use literacy skills weekly, IALS and PIAAC



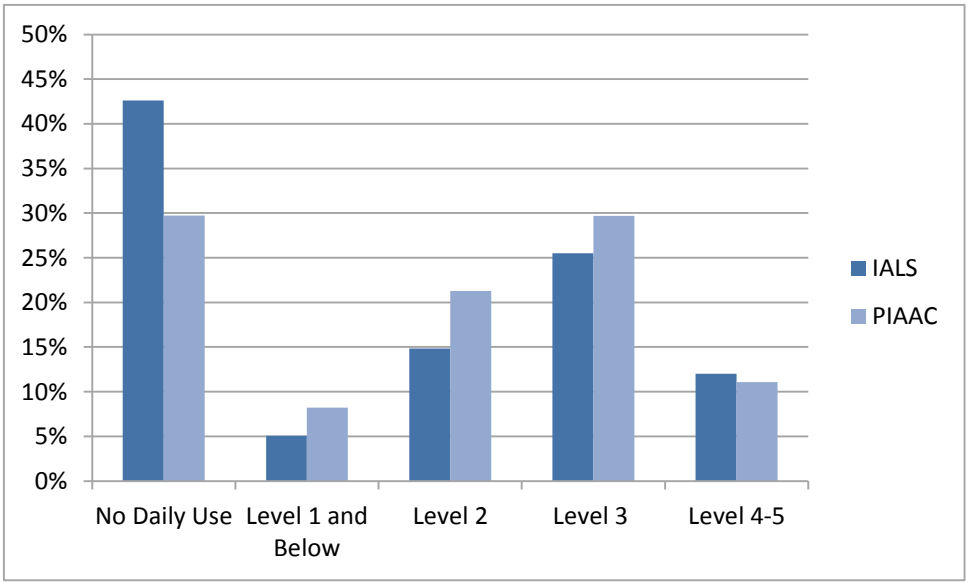
Source: International Adult Literacy Survey (1994-1998), and Survey of Adult Skills (PIAAC) (2012, 2015), Table A2.7.

There are relatively few significant differences in the changes in rates of use across countries (Appendix Tables A2.6 and A2.7). Germany and Italy both show decreases in the rates of weekly use for workers at all skill levels, with statistically significant differences from the country averages. Chile, Denmark, and the United States show decreases in weekly use at one or two proficiency levels, most of

which are significantly different from the average increase across countries. In addition, Ireland and Norway both show significantly faster increases in the rates of use for several proficiency levels for both daily and weekly use.

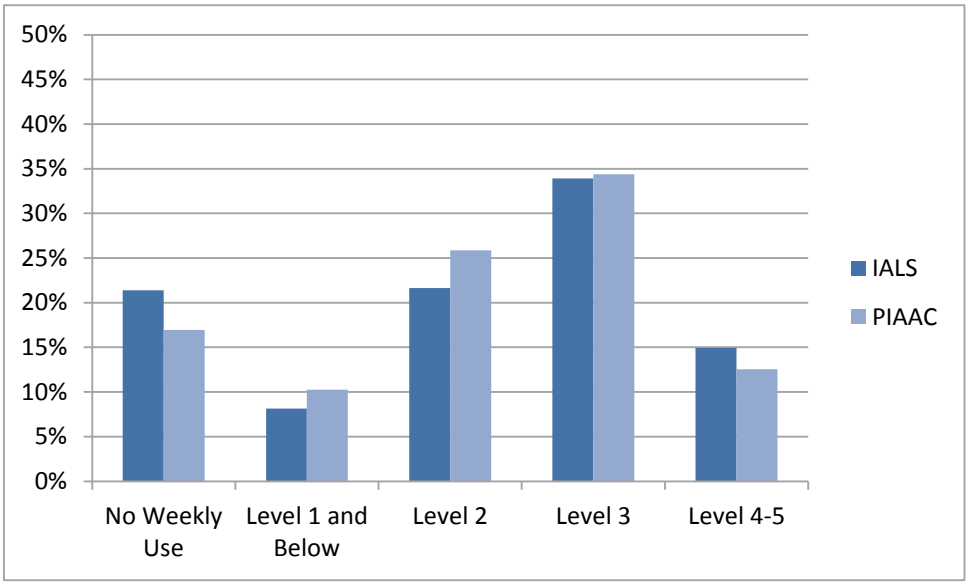
By combining the separate results on literacy proficiency and on literacy skill use, it is possible to see the change in the distribution of utilized literacy proficiency between IALS and PIAAC. Figure 2.8 shows the distribution of workers with respect to literacy proficiency and daily skill use, averaged across 18 OECD countries and economies included in both surveys. This figure shows that the largest increase in workers using literacy on a daily basis between the two surveys was for workers at Level 2 proficiency, along with smaller increases those at Level 1 and below and at Level 3. Figure 2.9 shows the change in the distribution of utilized literacy proficiency with respect to weekly rather than daily use. The largest increase in workers using literacy on a weekly basis is again for workers at Level 2 proficiency. The two figures show a modest decrease between the two surveys in the proportion of workers using literacy who are at Level 4-5 proficiency.

Figure 2.8 Distribution of workers by daily literacy use and level of proficiency, IALS and PIAAC



Source: International Adult Literacy Survey (1994-1998), and Survey of Adult Skills (PIAAC) (2012, 2015), Table A2.8.

Figure 2.9 Distribution of workers by weekly literacy use and level of proficiency, IALS and PIAAC



Source: International Adult Literacy Survey (1994-1998), and Survey of Adult Skills (PIAAC) (2012, 2015), Table A2.9.

The increase in the proportion of workers using literacy who are at Level 2 proficiency results from the combined effect of the increase in the proportion of workers at that level of literacy proficiency and the increase in the overall use of literacy skill. In contrast, the increases in the proportion of workers using literacy who are at Level 1 and Below and at Level 3 reflect only increased rates of use, since the proportion of workers at those skills was relatively stable between the two surveys. The decline in the proportion of workers using literacy who are at Level 4-5 results from the decline in the proportion of workers at those levels, which was only partly counterbalanced by the increased rates of use.

The increase in the proportion of workers using literacy who are at Level 2 is generally consistent across countries, with all countries showing an increase with respect to daily use and only three countries (Germany, Italy, and Slovenia) showing a decrease with respect to weekly use, none of which is statistically significant (Appendix Tables A2.8 and A2.9). However, the modest average decrease in the proportion of workers using literacy who are at Level 4-5 reflects clear differences in the patterns across countries. In contrast to the average decrease in workers using literacy who are at Level 4-5, Australia, Finland, the Netherlands, Poland and Slovenia show increases that are statistically significant for either daily or weekly use. In these countries, the increase in the proportion of workers using literacy who are at Level 4-5 averages 5 percentage points for daily use and 4 percentage points for weekly use, compared to the overall average decreases of 1 and 2 percentage points, respectively, for all included OECD countries and economies.

Comparison with the Economic Literature on Skill Use

The analysis of changes in skill and skill use presented in this chapter is rather different than the usual description of these changes in the economic literature. As noted in Chapter 1, the pattern often found for changes in workforce skill during the 1990s and 2000s is one of polarization, with increasing employment for workers with higher and lower skills and decreasing employment for workers with mid-level skills (e.g., Autor, 2015; Goos, Manning, and Salomons, 2014). In contrast, the distribution of literacy proficiency in the workforce has increased in the middle of the skill distribution and decreased at the upper level. Although workers of all skill levels are more likely to use their literacy at work, the net effect on the

portion of the workforce that regularly uses their literacy skills is an increase in regular users who have low to middle levels of literacy skill, with decreases both for workers who do not use literacy at all and for those who use literacy and have high levels of proficiency.

It is possible to reconcile these seemingly contradictory findings by understanding that they reflect two very different kinds of measures. The economic literature has used differences in wages across occupations as its primary measure of skill differences, along with analyses of education or task content by occupation. These three types of data provide information about work skills but that information is indirect, while also being quite broad, potentially reflecting the full range of important skills, as well as other labour market factors related to wages and tasks but not directly related to skill. In contrast, the literacy proficiency and skill use measures in PIAAC and IALS provide a direct measure of a single type of skill and questions related to the frequency of literacy use, but no measures of many other important skills and no questions about the amount of time, the importance or the complexity of the use of written materials. It is quite possible that skills other than general literacy or aspects of literacy use other than proficiency and frequency are driving the polarization findings in the economic literature.

The essential point is to be careful in understanding the measures that are being used and drawing conclusions from those measures that go beyond what the measures describe. Historically, the economics literature has had access to very little data directly related to worker skills - and as a result conclusions related to "skill" that are actually based on measures of educational attainment or wages or occupational clusters may not necessarily reflect understandings of skill that are meaningful to the education community. With respect to skill itself, it is important to consider the very different picture offered by PIAAC of the changes that have occurred over the past two decades - a picture that suggests an increase in the prevalence and use of mid-level skills in the workforce.

REFERENCES

Autor, D.H., 2015, Why Are There Still So Many Jobs? The History and Future of Workplace Automation, *Journal of Economic Perspectives*, *29*, 3, 3-30.

Goos, M., A. Manning, and A. Salomons, 2014, Explaining Job Polarization: Routine-Biased Technological Change and Offshoring, *American Economic Review*, *104(8)*, 2509-2526.

Hanushek, E.A., G. Schwerdt, S. Wiederhold, and L. Woessmann, 2015, Returns to Skills Around the World: Evidence from PIAAC, *European Economic Review*, *73*, 103-130.

OECD, 2016a, *Skills Matter: Further Results from the Survey of Adult Skills*, OECD Skill Studies, OECD Publishing, Paris.

OECD, 2016b, *The Survey of Adult Skills: Reader's Companion, Second Edition*, OECD Skills Studies, OECD Publishing, Paris.

OECD, 2013, *Technical Report of the Survey of Adult Skills (PIAAC)*, OECD.

Paccagnella, M., 2016, Literacy and Numeracy Proficiency in IALS, ALL and PIAAC, *OECD Education Working Papers*, No. 142, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jlpq7qglx5g-en.

Quintini, G., 2016, *Skills Use at Work / Why Does it Matter and What Influences It?*, OECD Publishing.

Quintini, G., 2011, Over-qualified or under-skilled: A review of existing literature, OECD Social, Employment and Migration Working Papers 121, OECD Publishing.

3 – METHODOLOGY FOR ASSESSING COMPUTER CAPABILITIES USING PIAAC

The Survey of Adult Skills (PIAAC) measures a set of general cognitive skills that are developed during formal education and widely used at work. The survey includes tests of skills in literacy, numeracy and problem solving with computers. To provide a way of anticipating the changes that technology may bring to the use of these skills in the future, the OECD asked a group of computer scientists to assess the capabilities of computers for answering the questions in the three skill areas included in the survey.

After first reviewing the goal for assessing computer capabilities, this chapter describes the participating experts and the series of questions they addressed to agree on a methodology for evaluating the PIAAC test questions. The common approach developed after extensive discussion involved providing a rating of the ability of current computer techniques to answer each test question after a one-year development period costing no more than USD 1 million, and using the same visual materials that were used by the adults who took the test. The rating options were Yes, No and Maybe, with respect to the capabilities of computers to answer each question. The chapter also summarizes the group's suggestions for improving the assessment approach in the future. The analyses of the ratings that resulted from this exploratory work are discussed in Chapter 4.

Objective for the exploratory assessment of computer capabilities

The goal for the exploratory assessment was to develop a way of obtaining information about computer capabilities in a form that would be meaningful to educators and education researchers. Educators and education researchers are usually familiar with the types of skills assessed on tests like the Survey of Adult Skills. They are also familiar with the ways those skills are developed in education and potentially used at work and in daily life. PIAAC was specifically designed to provide this type of information across countries, and other tests also provide such information for particular types of skills and particular groups of individuals. However, educators and education researchers usually have little familiarity with the kinds of capabilities currently being demonstrated by computer science. This lack of understanding makes it difficult for the education community to understand the kinds of changes computers are likely to bring to work and skill demand over the next several decades.

The OECD's analysis of computer capabilities was carried out to provide a way to help the education community begin to think about the ways that computers are likely to change the skill requirements for future jobs. If computers have demonstrated some of the general cognitive capabilities assessed by PIAAC, then it is likely that employers will begin to use that technology to perform some of the tasks requiring general cognitive skills. This will ultimately shift workers to a different set of tasks, resulting in job destruction, creation and transformation. The shift is likely to take place slowly, probably over a decade or more. However, it is useful for the education system to anticipate these changes since schools often help students acquire skills precisely because those skills are believed to be useful one or more decades in the future. If technology is likely to substantially change the work skills that will be useful in several decades, then the education community needs to begin to anticipate that change.

In addition to providing information to the education community, the exploratory work was also carried out to develop a more credible approach to assessing the capabilities of computers than has been done to date within economics. PIAAC allows an assessment of computer capabilities at a much more specific level of detail than has been available in the prior work discussed in Chapter 1. This prior work has involved general descriptions of occupations or occupational tasks that are too coarse for computer scientists to be able to understand exactly what behaviour is included. As a result, when such experts provide judgments about whether or not computers can carry out these tasks, it is generally not clear

exactly what tasks they have in mind and it is almost certain that different experts are thinking of different tasks when responding to the same descriptions. For example, a task description such as "reads reports" could be used in describing many different occupations and the difficulty of the relevant task could vary widely between and within occupations. A computer scientist responding to such a description could have many different possible tasks in mind when considering possible computer performance. In contrast, the PIAAC test questions involve precisely-defined tasks that allow computer scientists to analyse in detail the specific information provided and the necessary information processing to answer a specific question. The PIAAC test questions provide a much more credible basis for assessing the capabilities of computers, just as they provide a more credible basis for assessing the skills of adults than simply asking adults whether they are able to "read reports."

The basic plan to assess the capabilities of computers using PIAAC was to ask a group of computer scientists to review the test questions in PIAAC's three skill areas and identify the questions that could be answered by machines today. The expectation was that computer scientists who work in areas related to language understanding and reasoning would be able to make these judgments based on their expertise about the capabilities and limitations of existing techniques. Their assessments would then be used to help educators and education researchers understand the capabilities of computers with respect to these three general cognitive skills and to help economists develop a comprehensive programme for credibly assessing computer capabilities across the full range of work skills.

The assessment of computer capabilities was approached as an initial exploratory effort, with an expectation that it would take several additional attempts to refine a methodology for comparing machine capabilities to human skills. A relevant comparison is that it took several decades to develop and refine the approaches for comparing the skills of diverse individuals, including people from different cultures, people who speak different languages, and people with disabilities (e.g., National Research Council, 2002, 2004). With each of these expansions in the group of tested individuals, it was necessary to think carefully about which skills were being tested and why. When an existing test is given to a new group, it often becomes clear that some questions are unexpectedly hard - or easy - for the new group for reasons that have nothing to do with the skill being assessed. For example, a test of arithmetic may be difficult for a non-native speaker because of the language used on the test to give the instructions and describe the problems, rather than because of the mathematical difficulty of the problems. It is reasonable to expect similar problems when using tests to compare machine capabilities with human skills and it may take time to develop appropriate ways of addressing them.

Because of the expectation that this assessment of computer capabilities would be only the first step in the development of an approach to regularly monitor the growing capabilities of computers, the lessons about the approach for conducting the assessment are as important at this stage as the findings about the level of current computer capabilities.

Identifying a group of computer scientists

Over a period of ten months, approximately 60 computer scientists were contacted to provide input to the project. Initial recommendations for computer scientists were obtained from a set of social scientists who study the effects of computers on the labour market. These initial contacts were used to generate additional suggestions and the process was repeated until a full set of computer scientists had been identified who had appropriate expertise and were willing to participate in the evaluation.

Based on the initial set of contacts, the project identified a number of relevant areas of computer science for the assessment, including natural language processing, reasoning, common sense knowledge, computer vision, machine learning and integrated systems. The project attempted to find participants in each of these areas and was successful in identifying a small group of prominent experts who could

authoritatively address these different areas of computer science and were willing to participate in the exploratory work.[11] Table 3.1 lists the 11 participating computer scientists along with their areas of expertise.[12]

Table 3.1 Computer scientists providing assessments of computer capabilities

Computer Scientists	Expertise
Jill Burstein, Research Director, Natural Language Processing Group, ETS Research Division	Natural language processing, automated essay scoring, discourse analysis, educational technology
Ernest Davis, Professor of Computer Science, Courant Institute, New York University	Representation of common sense knowledge
Kenneth D. Forbus, Walter P. Murphy Professor of Computer Science and Professor of Education, Northwestern University	Qualitative reasoning, analogical reasoning and learning, spatial reasoning, sketch understanding, natural language understanding, cognitive architecture, reasoning system design, intelligent educational software
Arthur C. Graesser, Professor, Department of Psychology and Institute for Intelligent Systems, University of Memphis	Question asking and answering, text comprehension, inference generation, artificial intelligence, computational linguistics, discourse technologies, human-computer interaction, problem solving
Jerry R. Hobbs, Research Professor, Fellow and Chief Scientist for Natural Language Processing, Information Sciences Institute, University of Southern California	Computational linguistics, discourse analysis, artificial intelligence, parsing, syntax, semantic interpretation, information extraction, knowledge representation, encoding common sense knowledge
Rebecca J. Passonneau, Director, Center for Computational Learning Systems, and Senior Research Scientist, Columbia University	Computational linguistics, computational semantics and pragmatics, discourse analysis, data mining, methodology
Vasile Rus, Professor, Department of Computer Science and Institute for Intelligent Systems, University of Memphis	Artificial intelligence, machine learning, computational linguistics, automated and human question answering and asking
Vijay Saraswat, Research Staff Member and Manager, IBM TJ Watson Research Center	Cognitive computing, theoretical computer science, programming systems, artificial intelligence, natural language processing, machine learning, probabilistic logic
Jim Spohrer, Director, Global University Programs and Cognitive Systems Group, IBM	Artificial intelligence, cognitive systems for holistic service systems
Mark Steedman, Professor of Cognitive Science, School of Informatics, University of Edinburgh	Computational linguistics, artificial intelligence, cognitive science, speech generation, communicative use of gesture, parsing, semantics
Moshe Vardi, George Distinguished Service Professor in Computational Engineering and Director of the Ken Kennedy Institute for Information Technology, Rice University	Database systems, computational-complexity theory, multi-agent systems, design specification and verification

[11] The recruiting process also specifically attempted to identify a geographically balanced set of experts to ensure that a broad mix of research traditions from different countries would be reflected in the discussions. Although the project failed to find experts from a broad range of countries who were willing and able to participate, the experts who did participate in the meeting were well aware of work being carried out in different countries since computer science research is conducted on an international basis.

[12] In addition to the 11 computer scientists, the meeting included four social scientists familiar with applications of computers in the workplace: Charles Fadel, Center for Curriculum Redesign; Michael J. Handel, Northeastern University; Frank Levy, MIT and Harvard Medical School; and Alistair Nolan, OECD.

Structure of the assessment of computer capabilities

The assessment was carried out during a two-day meeting, with materials provided to the participants to review in advance. All participants were given copies of the test questions in all three skill areas - a total of 128 questions across the areas of literacy, numeracy and problem solving using computers.[13] The advance instructions and the initial discussion at the meeting addressed four primary issues about how to structure the task of evaluating the ability of computers to answer the test questions: 1) whether to assess individual questions or to use cut-points across the full set of questions; 2) whether to assess computer capabilities in the past and the future; 3) how much development work to allow for applying the computer techniques to the specific context of the test questions; and 4) how to address the extensive visual input used on the test. Each of these issues is discussed in turn below.

Rating individual test questions or using cut-points across the full set of questions

The copies of the test questions were grouped separately by the three skill areas and arranged in order of increasing difficulty for adults, using the difficulty score for each question that is calculated as part of the analysis and scaling of the test results. The advance instructions suggested that the participants should identify cut-points in the series of questions (from easy to difficult) between the questions that could be answered by computers now and those that could not. Using cut-points was suggested to provide an easy way of aggregating and comparing the ratings of the different experts at the meeting. Recognizing that the order of difficulty of the questions would not necessarily be the same for computers as for people, the instructions suggested that the participants identify questions that were seriously out of order with respect to their difficulty for computers so that they could be considered separately.

In the discussion at the meeting itself, however, it became clear that the approach using cut-points did not work for the participants. Only about half of them had been able to evaluate the questions using cut-points and most had strong practical and theoretical objections to the approach. The group certainly recognized that there are some ways in which problems that are more difficult for people will also be more difficult for machines: for instance, because they involve longer texts, require more inferences and include more possible wrong answers that need to be avoided. However, they also noted a number of ways that the difficulty of the questions is substantially different for people and machines. On the one hand, questions are often difficult for people if they involve long repetitive texts or complicated calculations, factors that often pose little difficulty for machines. On the other hand, many questions that are easy for people involve interpreting pictures or social contexts, or coordinating information from pictures and text, factors that are often quite difficult for computers. Because of these arguments, the group decided it would be better to rate computer capabilities with respect to each question. While potentially being more time-consuming, this approach avoided the necessity of making an assumption about how the ordering of difficulty of the test questions for computers relates to the ordering of their difficulty for people.

Giving ratings for the past and the future

The advance instructions asked the computer scientists to make their initial assessments with respect to the current capabilities of computers in 2016. Although there is great interest in the likely future capabilities of computers, the goal of the assessment was to avoid speculation on the initial rating and to assess computer capabilities in a way that could be justified by results demonstrated in the published

[13] See Chapter 2 for a brief description of the survey administration and Chapter 4 for a brief description of the different skill areas. PIAAC is usually administered on a computer when data are collected from adults, but the assessment of computer capabilities was carried out using static screen shots. As a result, in some cases only part of the question was available for evaluation. For more information on the design, administration and results of the Survey of Adult Skills see OECD (2012, 2016a, 2016b).

research literature. After the initial assessment, the experts were also asked to consider how their assessments would have been different in 2006 and how they might be different in 2026. The point of introducing these alternative dates was to provide a way of thinking about the change in capabilities over time. Because of the speculation involved with projecting improvements to 2026, the initial plan was to rely primarily on the ratings for 2006 to look at change over time, since these ratings could be linked to the published literature and thereby avoid speculation.

It turned out, however, that the ratings for 2006 were difficult for the group to provide. Several of the experts explained this as being related to the difficulty of trying to imagine not knowing something that you already know: the improvements that have taken place since 2006 have been fully integrated into expert thinking and it is hard to identify when particular changes took place without going back to reconstruct the developments from the published literature itself.[14]

The group found it easier to think about likely improvements by 2026, while acknowledging that these projections could be quite wrong. There is a long history in AI of wildly optimistic projections of success in resolving problems that turned out to be much more difficult than was originally believed.[15] Three experts provided a complete set of projections for the test questions for 2026. Although the group was generally more comfortable in projecting forwards than backwards, one expert pointed out that a projection of 5 years to 2021 would be more natural because many grant applications require investigators to project the results of their own research over a 3-to-5-year period; as a result, researchers have regular experience in estimating the degree of change that can occur over this shorter period.

Considering the allowed effort to develop computer systems for the test questions

It was necessary for the computer scientists to consider how much development work would be allowed when thinking about the ways that current computer techniques might need to be adapted to the context of the test questions in the three skill areas. Although the questions are designed to be familiar to the general adult population, there is no reason to expect that existing computer systems would have already been developed for the types of questions included on the test.

Some computer techniques - like text search - can be applied to many different contexts without special preparation, but other computer techniques need to be adapted to specific contexts. This adaptation can involve training the system on a set of relevant examples or coding information about specific vocabularies, relationships or types of knowledge representation, such as charts and tables. In asking the experts to consider the possibility of developing a computer system using current techniques to answer the test questions, it was necessary to put some boundaries on the size of the hypothetical development effort that would be required.

Two rough criteria were used in selecting an appropriate set of boundaries for the development effort that the experts should have in mind when making their judgments. First, the assessment was intended to reflect the application of current computer techniques, not the creation of completely new computer techniques. If a development effort uses large quantities of people, time and funding, it looks more like a

[14] In fact, only one expert provided a complete set of ratings for 2006, although two others categorized the difficulty of the problems and suggested a way that their categories might relate to capabilities in 2006.

[15] One example cited in the discussion was the case of computer vision, which was proposed as a summer research project in the mid-1960s and now a half century later is still one of the hardest problems in AI (Papert, 1966). However, unexpected successes from new techniques can also lead to the opposite result. For example, after the recent victory of Google DeepMind's AlphaGo program over one of the world champions of the game of Go, some experts commented that such success was not anticipated for at least another decade (Silver et al., 2016).

research effort to develop new techniques than a development effort to apply current techniques. Second, the rating was intended to reflect the level of investment that a large company might be willing to invest to automate a particular task that is frequently performed within the organization. In this sense, the test questions were being used as a proxy for company-specific or job-specific tasks using general cognitive skills that a company might consider automating and might therefore be willing to invest in developing. Both of these criteria suggest a relatively limited development effort.

The advance instructions suggested that the computer scientists should think about a development effort representing roughly the work that could be done by a few people during a single year. During the discussion at the meeting, this constraint was further specified to involve an expenditure of no more than USD1 million for development.

How to approach the use of visual materials

The Survey of Adult Skills uses materials in its test questions that are similar to the types of written materials that adults encounter at work and in their everyday lives. These materials include signs, labels, advertisements, charts, tables, webpages, maps, drawings and photographs (OECD, 2016a). This range of test material is substantially different than more academic tests that might assess literacy only with narrative texts, and numeracy only with mathematical problems.

The diverse range of material used in the Survey of Adult Skills raises challenges for computers and the group spent substantial time figuring out how to address those challenges. In many cases, the diversity of input is included precisely because of the desire to assess whether adults are able to use information from such different sources. Most of the different types of materials are in general use and it is reasonable to assume that most adults will have been exposed to similar materials at school, at work or in their daily lives.

In other cases, however, the diversity was likely included to produce material that looks realistic, such as advertising with colourful designs and writing in distinctive layout. In these cases the extra realistic features probably do not cause any extra difficulty for the adults who take the test; indeed, extra realism may well make the materials more familiar to many adults and easier to use. However, such materials can make the questions substantially more difficult for computers. One example that the group discussed extensively was the easiest numeracy question for people, which uses a photograph of two packages of bottled water and asks how many bottles are in the packages. The numeracy aspect of the problem involves a simple multiplication, which is why the problem is so easy for people. However, the visual interpretation needed to answer the question - which people also find easy - is quite difficult for machines because the packaging makes many of the bottles hard to see. This question received the lowest average rating for computers across the group of experts because of the difficulty machines have in interpreting this sort of image by combining the image itself with the right knowledge about the physical world.

The group discussed two options for addressing the visual material in the test questions. The first option involved assuming that the visual input would be transformed into a textual or numerical form, such as extracting the written material from an advertisement or turning a graphical chart into a digital table. In this option, a computer would answer the question using transformed materials that eliminate the problem of interpreting the visual input. The second option involved taking the visual input as given, requiring the computer to solve the same visual interpretation problem that people need to solve. The group decided to adopt the second option to preserve the integrity of the full set of test questions. As a result, some of the questions that are identified as ones that computers could not answer - like the easiest question in numeracy discussed above- were identified as too difficult for computers primarily because they use visual material that is hard for computers to interpret.

Carrying out the second option added an extra practical difficulty for the exploratory assessment: because visual processing is often not considered relevant to work in computer language and reasoning, many of the participating computer scientists did not have extensive knowledge about current capabilities in vision. To make up for this, the participating experts who do have some knowledge of those capabilities discussed the visual features that were likely to be easy or difficult in a sample of the problems. As a result, the judgments about the difficulty of the visual aspects of the questions reflected a more limited range of expertise across the group than the judgments about the language and reasoning aspects of the questions.

Final specifications of the assessment exercise carried out at the meeting

After extensive discussions on the first day of the meeting, which involved hearing from each of the computer scientists about the way they had approached the assessment task in preparing for the meeting, the group agreed on a common approach for conducting the assessment. This common approach involved providing a rating of the ability of current computer techniques to answer each test question after a one-year development period costing no more than USD 1 million, and using the same visual materials that were used by the adults who took the test. The rating options were Yes, No and Maybe, with respect to the capabilities of computers to answer each question. All 11 computer scientists in the group provided these ratings for the literacy and numeracy questions. In addition, six of the experts provided ratings for the third skill area of problem solving using computers, and three of them provided ratings for computer capabilities in 2026. The assessment ratings are analysed in Chapter 4.

Suggestions to improve the approach to assessing computer capabilities

The issues discussed above shaped the group's decision about how to provide a comparable set of assessments of computer capabilities for answering the test questions. In addition, the participants made a number of other suggestions for assessing computer capabilities that they did not have time to pursue at the meeting. The common theme linking these different suggestions was the possibility of finding ways to resolve disagreements in the ratings across the group. The section that follows discusses two types of suggestion: one focusing on improving understanding of the test questions and the other focusing on improving understanding of the capabilities of current techniques.

Improving understanding of the test questions

As the group discussed different questions at the meetings, there were a number of cases where participants realized they had misunderstood the requirements of a particular question. Sometimes this realization led them to decide that the question was actually easier or more difficult for computers than they had originally thought. For example, the instructions for a number of questions say that the test-taker should highlight the passages in the text that provide an answer to the question, rather than directly provide the answer itself. In some cases, this difference - between highlighting the relevant text and independently specifying the answer - significantly affects the difficulty of providing an answer. Sometimes some of the participants had missed this distinction in their evaluation and the discussion ended up bringing the group closer to consensus about the difficulty of the question for computers.

To help make the rating process more systematic, several participants suggested it should be carried out in two stages, first identifying the different types of capabilities needed for each problem and then identifying what computers can do in each area. For example, the group's extensive discussion of the challenges raised by the visual materials used in some of the questions showed the importance of identifying the questions that require visual interpretation. The group discussed some key contrasts in visual processing requirements - such as the difference between black-and-white and colour images - which are related to limits in current computer capabilities and could be used to code specific aspects of

the visual materials used in the questions. Although a two-stage method seemed like a promising way to approach the rating process systematically, the group did not have enough time to apply it. Clearly the second-stage assessment requiring multiple judgments for each test question would be more time-consuming than the single judgments the experts made at the meeting. In addition, several of the experts thought it would be time-consuming in the first stage to agree on a set of categories to describe the different types of capabilities.

One concern that was raised during the discussion was that tests generally focus on assessing capabilities that are hard for people and often omit capabilities that are generally easy for people but hard for machines, such as vision and social interaction. While this makes sense for a test used for people, it raises problems in interpreting computer performance on the test. If a test omits capabilities that most people share but machines do not, then the results would overestimate computer performance in situations where those capabilities are important. On the other hand, if a test includes such capabilities, then computers may perform poorly primarily because of those capabilities, rather than because they lack the primary capabilities being assessed. In this case, the results would underestimate computer performance in situations where these sorts of capabilities are not important. Without being aware of the potential confounding role of the capabilities that are generally easy for people, it can be misleading to use estimates of computer capabilities from human tests to draw conclusions about the types of work tasks that computers might be able to perform.

This issue of the inclusion of additional capabilities was illustrated during the meeting by the extensive discussion of the visual materials used in PIAAC, most notably with the example of the easiest numeracy question (discussed above in the section on visual materials). This question is clearly easy for most adults and the numerical reasoning aspect of the question is also easy for machines. However, the group gave this question the lowest rating with respect to computer capabilities because of the difficulty machines would have interpreting the photograph it uses of packaged water bottles. This question provides a good measure of computer numeracy capabilities in combination with visual interpretation, but a misleading measure of computer numeracy capabilities on their own.

In general, the experts felt that the diverse material used in PIAAC does a better job representing capabilities that are easy for people but difficult for computers than is the case for many narrow academic tests. However, it would be useful to analyse the questions separately that require these additional skills from the questions that do not. This more precise analysis of the test questions would make it easier to understand where low computer performance is related specifically to the primary skills that are being tested by the Survey of Adult Skills - literacy, numeracy and problem solving - and where that performance is related to the need for additional capabilities such as vision.

Some additional capabilities, such as social interaction, are not reflected at all in the Survey of Adult Skills. For such capabilities, there are no relevant questions on the test that could be identified by a detailed analysis of the test questions. It would be helpful to simply identify that these skills have been omitted from the test and take that limitation into account when using the assessment results to analyse the potential effects of computers in different work settings. For example, an assessment of computer capabilities in literacy using PIAAC will probably be more useful in analysing the automation potential of language-related tasks in administrative jobs than in customer service jobs, because social interaction is more important for the latter. Another option would be to use other tests to assess these additional capabilities.

Finally, another question raised by one of the participants concerned how to generalize the skills being measured on the test and therefore how to evaluate the underlying computer capabilities. When the computer scientists considered whether a particular question could be answered, they were interested in proposing general computer techniques that could potentially be successful on a wide range of comparable

questions, rather than techniques geared specifically to work on a single question. But it was sometimes difficult to know what questions would be truly comparable, since small differences in wording can often make a question much harder or easier for people, and presumably for computers as well. One way to address the range of generalization of the skills being tested would be to provide more examples of test questions. Although this is not possible with PIAAC, which has a limited set of questions, many other standardized tests have large sets of practice questions that illustrate the range of material that will be tested.

Improving understanding of computer capabilities

There was general agreement across the group that their expertise was weak in the areas of computer vision and machine learning. Although there were participants who were familiar with work in each of these areas, the group did not include researchers whose primary work is focused on one of these areas. The participants strongly recommended that future work to assess computer capabilities with respect to the Survey of Adult Skills include researchers with these specialties.

Throughout the meeting discussion, there were numerous exchanges about the level of performance achieved by particular computer techniques. In most cases, all of the computer scientists were generally aware of the techniques mentioned, but not all of them knew about particular recent results or details about how a technique had been applied. Given the limited time at the meeting, usually the exchanges about the details of a specific technique were limited to mentioning a relevant research article. Unlike the exchanges during the meeting about the nature of the questions, the exchanges that occurred about the performance of particular techniques did not appear to cause any of the experts to re-evaluate their conclusions about the difficulty of some of the test questions, except in the area of computer vision. With respect to computer techniques used for language and reasoning, it appeared that the group would have required substantially more time to discuss their different perspectives on particular techniques in order to move closer to a consensus in their assessments.

One question raised by the discussion was what conclusions to draw at this exploratory stage from the disagreements in the assessments made during the meeting, given the limited time to discuss the papers that the participants cited as justification for their conclusions. If only one person in the group knew about a new technique that they believed would allow computers to be successful on a particular type of question, then without the time to share the details across the group most of the participants would say that computers could not be successful. The benefit of working toward group consensus is that it allows the possibility that the one person who knows about a new technique has the opportunity to educate everyone else. Of course, this can also go the other way, with a single sceptic who understands the limitations of a particular technique convincing everyone else that it would not be successful on a particular type of question. Because of the lack of time to work towards a full consensus understanding of the different computer capabilities, the analysis of the assessment ratings in Chapter 4 uses a variety of approaches to explore the range of views across the group.

Finally, several of the participants argued that discussion and analysis alone would ultimately be insufficient for reaching a consensus about the ability of current techniques to answer the test questions. Instead, these experts suggested that it would be necessary in some cases to actually apply computer techniques to the test questions to see whether they would be successful. Such tests have frequently been done in the field of computer science by holding competitions, which can sometimes attract substantial interest (e.g., Quillen, 2012; Visser and Burkhard, 2007). However, for resolving questions about the potential performance of particular techniques, it could also be effective to commission specific research groups who work with those techniques to apply them to a set of questions to assess their performance.

Summary of possible extensions for future work

The discussions at the meeting produced a range of suggestions for deepening the assessment of computer capabilities on a set of tested skills. With respect to the test questions themselves, the meeting discussion suggested three possible extensions for future work: 1) conducting a two-stage evaluation with separate analyses of question requirements and computer capabilities; 2) considering the full set of work skills and identifying skills that are omitted from the test but that may be important in some work contexts where the tested skills are used; and 3) working with tests with a larger number of example questions. With respect to the computer techniques, the meeting discussion suggested another three extensions: 4) expanding the range of computer science expertise included in the discussion; 5) reviewing a set of key research papers in greater detail; and 6) obtaining empirical results about the ability of computers to answer the test questions, particularly with respect to techniques or question types where the group was not able to reach consensus. These extensions provide a set of approaches that could be pursued in future work to sharpen the assessment ratings discussed in Chapter 4.

REFERENCES

National Research Council, 2002, *Methodological Advances in Cross-National Surveys of Educational Achievement*, Board on International Comparative Studies in Education, A.C. Porter and A. Gamoran, eds., Washington, DC: The National Academies Press.

National Research Council, 2004, *Keeping Score for All: The Effects of Inclusion and Accommodation Policies on Large-Scale Educational Assessments*, Committee on Participation of English Language Learners and Students with Disabilities in NAEP and Other Large-Scale Assessments, J.A. Koenig and L.F. Bachman, eds., Washington, DC: The National Academies Press.

OECD, 2016a, *Skills Matter: Further Results from the Survey of Adult Skills*, OECD Skill Studies, OECD Publishing, Paris.

OECD, 2016b, *The Survey of Adult Skills: Reader's Companion, Second Edition*, OECD Skills Studies, OECD Publishing, Paris.

OECD, 2012, *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*, OECD Publishing. Paris. http://dx.doi.org/10.1787/9789264128859-en.

Papert, S., 1966, The Summer Vision Project, Artificial Intelligence Group, Vision Memo. No. 100, Massachusetts Institute of Technology. Available at http://hdl.handle.net/1721.1/6125 [accessed 24 January 2017].

Quillen, I., 2012, Hewlett Automated-Essay-Grader Winners Announced, EdWeek, May 9. http://blogs.edweek.org/edweek/DigitalEducation/2012/05/essay_grader_winners_announced.html [accessed 24 January 2017].

Silver, D., A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcouglu, T. Graepel, and D. Hassabis, 2016, Mastering the Game of Go with Deep Neural Networks and Tree Search, *Nature*, 529, 484-489.

Visser, U., and H.-D. Burkhard, 2007, RoboCup: 10 Years of Achievements and Future Challenges, *AI Magazine, 28(2)*, 115-132.

4 – ASSESSMENT OF COMPUTER CAPABILITIES ON PIAAC TEST QUESTIONS

This chapter describes the results of the exploratory assessment of computer capabilities on the Survey of Adult Skills (PIAAC). The assessment was carried out by a group of computer scientists using the approach described in Chapter 3. Most of the attention in the assessment focused on the ability of current computer techniques to answer test questions in literacy and numeracy. In these two skill areas, all 11 participating computer scientists provided ratings for each question, using a similar approach. Each expert provided a rating of Yes, No or Maybe for the ability of current computer techniques to answer each test question after a one-year development period costing no more than USD 1 million, and using the same visual materials that are used by adults who take the test. In addition, six of the participants provided ratings for the third skill area of problem solving using computers, and three of the participants provided ratings for possible computer capabilities in 2026 for all three skill areas. This chapter discusses the results for the different skill areas in turn: literacy, numeracy and problem solving using computers.

In general, the experts projected a pattern of performance for computer capabilities in the middle of the adult proficiency distribution on PIAAC. In literacy, these preliminary results suggest that current computer techniques could perform roughly like adults at Level 2 and that Level 3 performance is close to being possible. In numeracy, the preliminary results suggest that computer performance is roughly at Level 2 and that Level 3 or 4 is close to being possible. In problem solving with computers, the preliminary results suggest that computer performance is roughly at Level 2 and that Level 3 is close to being possible.

Ratings of computer capabilities to answer the literacy questions

Literacy skill in PIAAC is defined as the "ability to understand, evaluate, use and engage with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential" (OECD, 2012). The test includes the decoding of written words and sentences, as well as the comprehension, interpretation, and evaluation of complex texts; it does not include writing. The test includes questions using different types of texts, including both print-based and digital texts, as well as both continuous prose and non-continuous document texts, and questions that mix several types of text or include multiple texts. The questions are drawn from several contexts that will be familiar to most adults in developed countries, including work, personal life, society and community, and education and training.

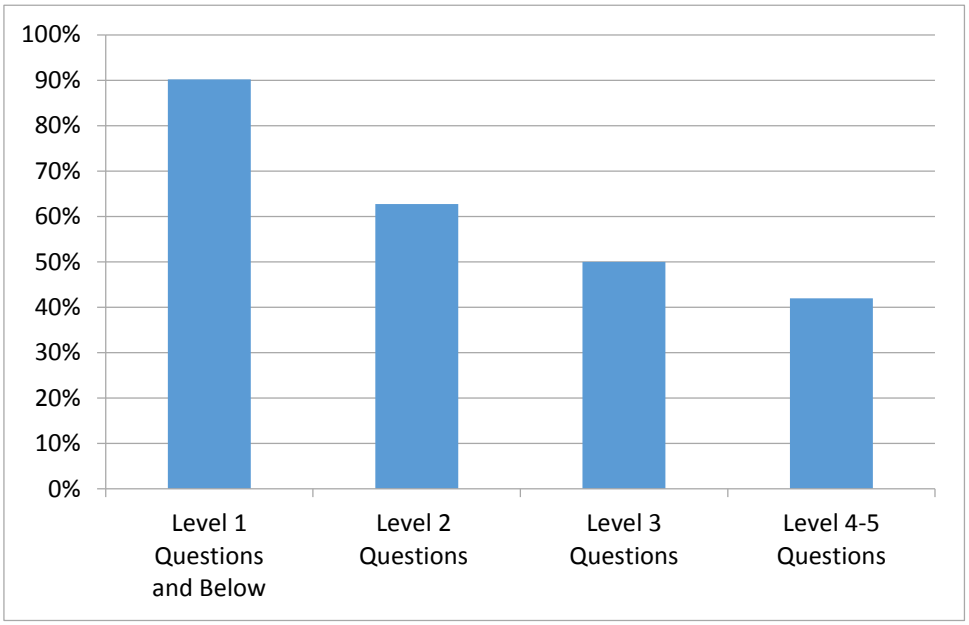
Literacy proficiency is described in terms of 6 proficiency levels, ranging from below Level 1 to Level 5. The easier test items involve short texts on familiar topics and questions that can be matched directly to a passage of text. The harder test items involve longer and sometimes multiple texts on less familiar topics, questions that require some inference from the text, and distracting information in the text that can lead to a wrong answer. For example, a Below Level 1 item has several brief paragraphs about a union election with a simple table showing the votes for three candidates and asks which candidate received the fewest votes. An example Level 2 item shows a simple website about a sporting event and asks for the phone number for the event organisers, which is not shown directly but can be found by following a link marked "Contact Us." And an example Level 4 item provides the result of a library search for books related to genetically modified foods and asks which book argues that the claims made both for and against genetically modified foods are unreliable. This item requires the test-taker to interpret the

information in the title and brief description for each book and to avoid many books that are superficially related to the question but not a correct response (OECD, 2013a).[16]

Computer literacy ratings by question difficulty

Figure 4.1 shows the average assessment ratings of computer capabilities on the questions at each literacy proficiency level.[17] For each question, the answers of the different experts are averaged together, counting a Yes as 100%, a Maybe as 50%, and a No as 0%. These average expert ratings by question are then averaged together for all questions in each proficiency level. The average expected performance ranges from a high of 90% on the questions that are easiest for adults (Level 1 and Below) to 41% on the questions that are most difficult for adults (Levels 4 and 5).

Figure 4.1 Expert ratings of computer capabilities to answer literacy questions, averaged with Maybe=50%, by level of PIAAC question difficulty



Source: Table A4.1.

Although the average expected performance for computers by proficiency level decreases as the questions become more difficult for adults, there are big differences across the different questions within each proficiency level. Overall, the correlation coefficient across the individual questions between the average expected rating for computers and the question difficulty score for adults is -0.61.

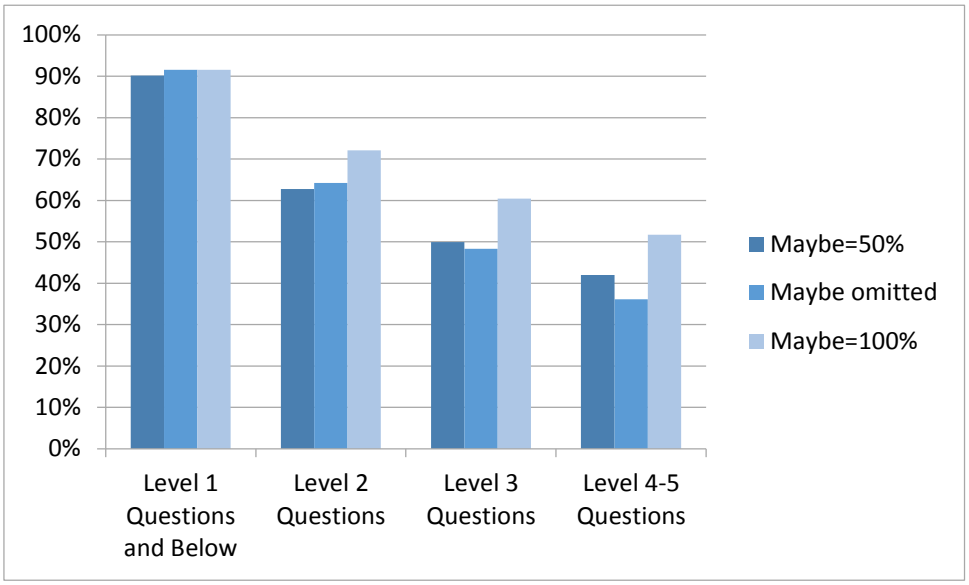
The participants discussed alternative meanings for their Maybe ratings, with some saying they used a Maybe rating to reflect genuine uncertainty about whether computers could answer a question and others saying they used Maybe when they believed computers could probably answer a question but were not completely sure. To reflect these two possible interpretations, Figure 4.2 provides two alternative

[16] More information about the Survey of Adult Skills and examples of the literacy questions are provided in OECD (2013a, 2013b). Full descriptions of the literacy proficiency levels are provided in the Appendix in Table B4.1.

[17] Complete assessment ratings for current computer capabilities by literacy question and expert are provided in the Appendix in Table B4.2.

averages, one that omits the Maybe ratings (to reflect genuine uncertainty) and one that groups them with the Yes ratings. The version with the Maybe ratings omitted from the averages produces little change in the overall results. The version with Maybe counted as 100%, like the Yes ratings, increases the expected performance on Levels 2-5 by about 10 percentage points each. It is not surprising that alternative codings produce relatively small differences since the Maybe rating was used in only 19% of the judgments.

Figure 4.2 Expert ratings of computer capabilities to answer PIAAC literacy questions, averaged with alternative coding of Maybe ratings, by level of question difficulty



Source: Table A4.2.

Accounting for differences in areas of expertise in the literacy ratings

The expected performance ratings shown in Figure 4.1 count the assessments of each of the computer scientists equally. As a result, a high score is possible on a particular question only if most of the computer scientists in the group know about a technique that could be used to successfully answer the question. This way of aggregating the results may be overly conservative in some cases since it effectively prevents new techniques that only a few of the experts know about from leading to an aggregate Yes rating. Although most of the experts will know about the well-established techniques that have been studied in the field for some time, each of the experts probably knows about a somewhat different set of results related to newer techniques. For those test questions that could potentially be answered by newer techniques but not older techniques, it may be that only a few of the experts in the group know about relevant research that would demonstrate this ability.

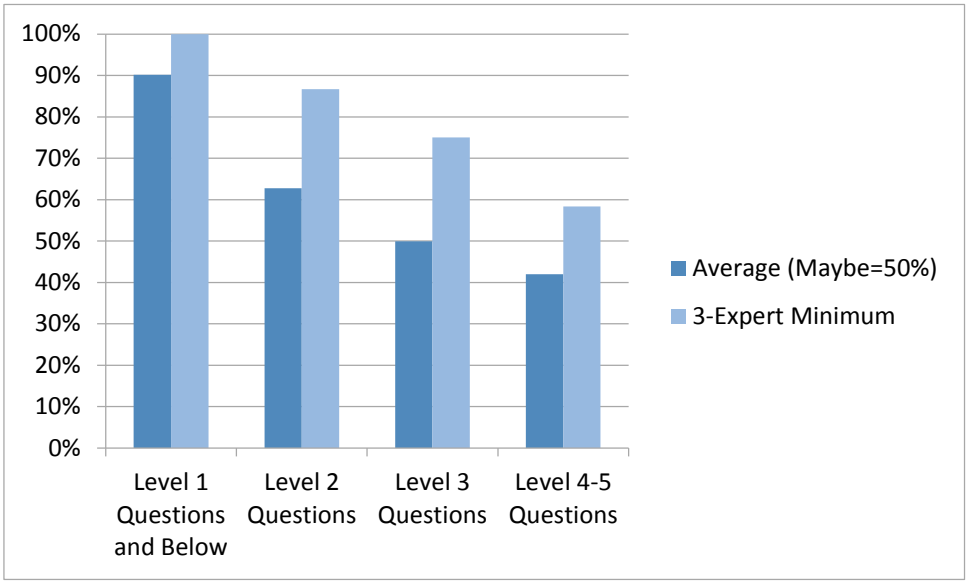
An alternative way of aggregating the ratings across the group would be to assign an aggregate Yes rating for computers if some minimum number of experts rates that question as Yes.[18] This approach takes into account the differences in techniques that the different experts in the group know about. If several experts know about a technique that could be used to answer a particular question, then it would be reasonable to count that as a question that computers are likely to be able to answer - even if the other

[18] Another approach to reflecting different levels of expertise with respect to the different questions would have been to allow the experts to give a rating on their confidence in their judgments for each of the questions. This approach was not discussed or used at the meeting, but could be explored in future work.

experts do not know about that technique and believe that computers could not answer the question successfully.

Figure 4.3 shows the results of an analysis using a 3-expert minimum where each question is counted as Yes if at least 3 of the 11 computer scientists rated it as a Yes. With this approach to aggregating the results, the proportion of questions that the group expects could be answered successfully by computers ranges from 100% of the easiest questions (Level 1 and Below) to 58% of the most difficult questions (Levels 4 and 5). These results suggest a substantially higher level of computer success on the questions than when the ratings are simply averaged across the group.

Figure 4.3 Expert ratings of computer capabilities to answer PIAAC literacy questions, comparing average using Maybe=50% and 3-expert minimum, by level of question difficulty

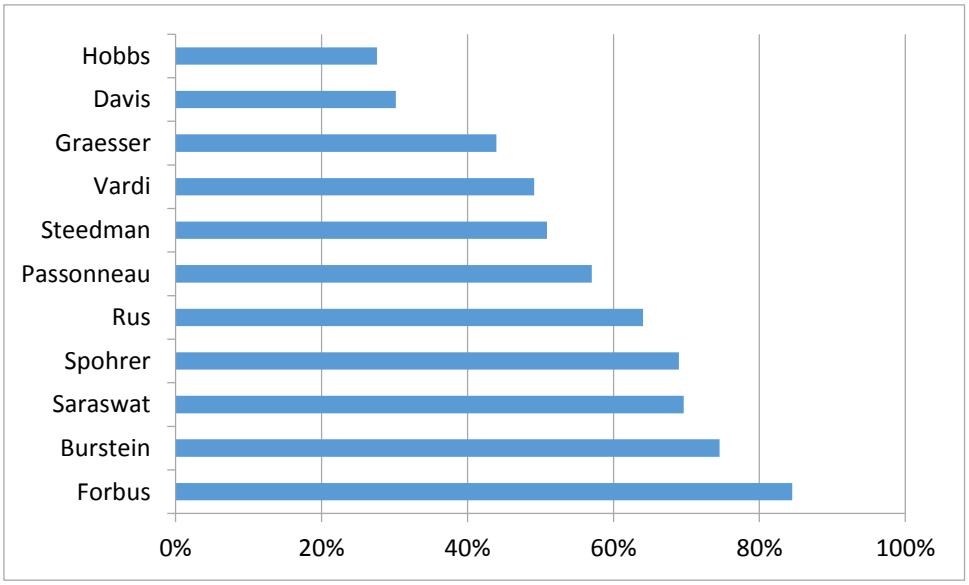


Source: Table A4.3.

Computer literacy ratings by expert

In addition to different types of expertise related to different computer techniques, the computer scientists in the group had different overall levels of optimism about the general ability of computers to answer the literacy test questions. To compare the level of optimism, Figure 4.4 shows the average rating across all literacy questions for each expert, counting a Yes as 100%, a Maybe as 50%, and a No as 0%. The scores range 56 percentage points across the experts, from 28% for Hobbs to 84% for Forbus. The average for the group is 56%. Changing the scoring for Maybe - to omit the rating from the average or to count it as 100% - does not make an appreciable difference to the range in the average rating across the experts.

Figure 4.4 Expert ratings of computer capabilities to answer PIAAC literacy questions, by expert



Source: Table A4.4.

The fact that some of the experts were much more optimistic than the others raises a question about the 3-expert minimum analysis in Figure 4.3 that counts a question as Yes if at least three experts give it a Yes. Rather than different types of expertise, this aggregation approach may simply reflect the judgments of the most optimistic experts in the group, since it would be possible for a question to receive a Yes with only the results of the three most optimistic experts (Forbus, Burstein and Saraswat). To adjust for this, one might add the additional requirement that at least one of the experts saying computers can answer the question is not in the group of the top three optimists. Adding this extra requirement does not substantially change the results, only modestly decreasing the computer rating for Level 3 questions from 75% to 67%, and the rating for Level 4 and 5 questions from 58% to 50%.

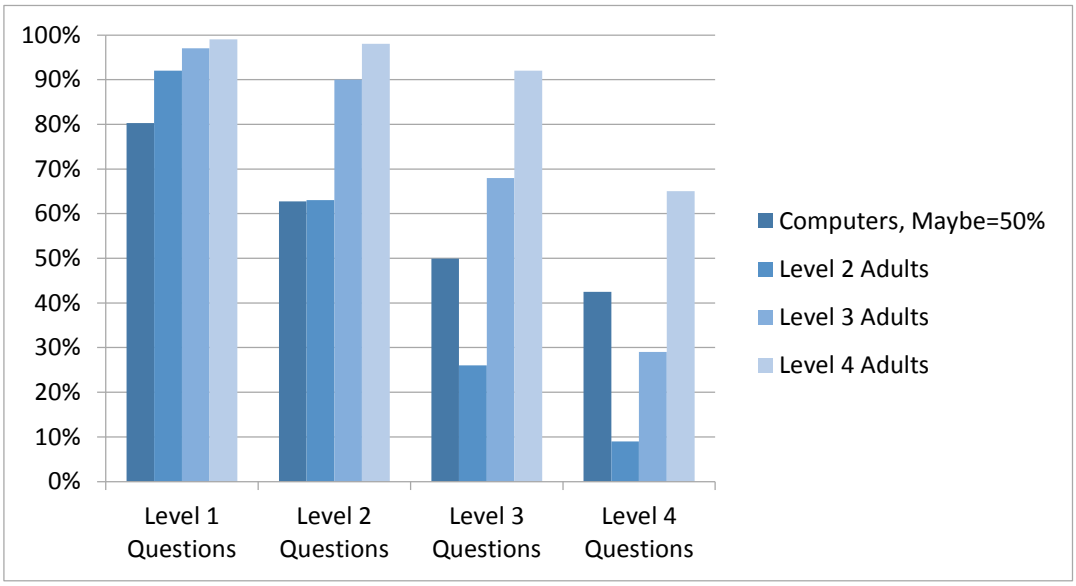
The different level of optimism across the group also raises the possibility of excluding the experts who are more extreme, focusing on those in the middle as representing a view that might be more representative of a consensus of the field. However, averaging the ratings across the five experts in the middle (Vardi, Steedman, Passonneau, Rus and Spohrer) produces results that are very close to the simple average for the full group.

Comparing the computer literacy ratings to human scores

The scoring process for the Survey of Adult Skills uses item response theory to calculate difficulty scores for each question as well as proficiency scores for each adult, with the scores for both questions and people placed on the same 500-point scale (OECD, 2013c). Each adult who takes the test is placed at the level where they answer two-thirds of the questions successfully. As a result, an adult with a literacy proficiency of Level 2 can successfully answer Level 2 questions about two-thirds of the time. Generally, people will be more successful in answering questions easier than their level and less successful answering questions harder than their level. For example, an average adult at the mid-point of Level 2 can answer 92% of Level 1 questions and only 26% of Level 3 questions (OECD, 2013b, Table 4.6).

Figure 4.5 compares the expected computer ratings for literacy with the performance of adults at three different levels of literacy proficiency, using the average of the expert ratings and coding Maybe as 50%.[19] Compared to Level 2 and 3 adults, the computer ratings show less change across the different levels of question difficulty, with lower expected performance on the easier questions than people show and higher expected performance on the harder questions. The computer ratings are worse than Level 2 adults on the Level 1 questions, match Level 2 adults on the Level 2 questions, and are substantially better than Level 2 adults on the Level 3 and 4 questions. On the Level 4 questions, the computer ratings are also above the Level 3 adults.

Figure 4.5 Comparing computer literacy ratings with adults of different proficiency, using average rating with Maybe=50%, by level of PIAAC question difficulty

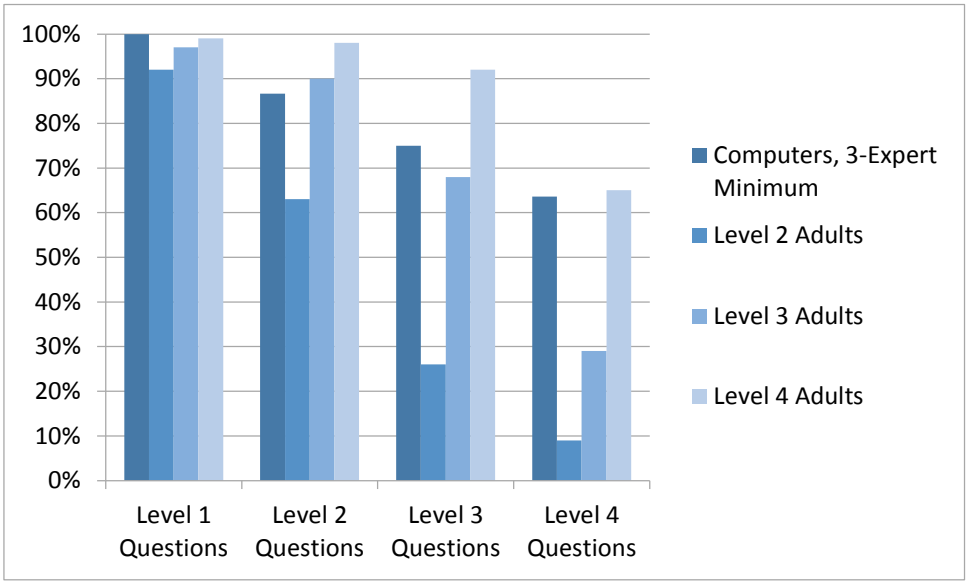


Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A4.5.

Figure 4.6 compares the expected computer ratings and adult performance using the 3-expert minimum analysis. With this alternative, the computer ratings are better than Level 2 adults for questions at all levels of difficulty. The computer ratings are better than Level 3 adults for questions at all levels of difficulty except for the questions at Level 2, where the computers are roughly comparable.

[19] The results are somewhat different than shown in Figure 4.1 for the expected computer ratings at the top and bottom because the questions below Level 1 and at Level 5 are excluded.

Figure 4.6 Comparing computer literacy ratings with adults of different proficiency, using 3-expert minimum, by level of PIAAC question difficulty



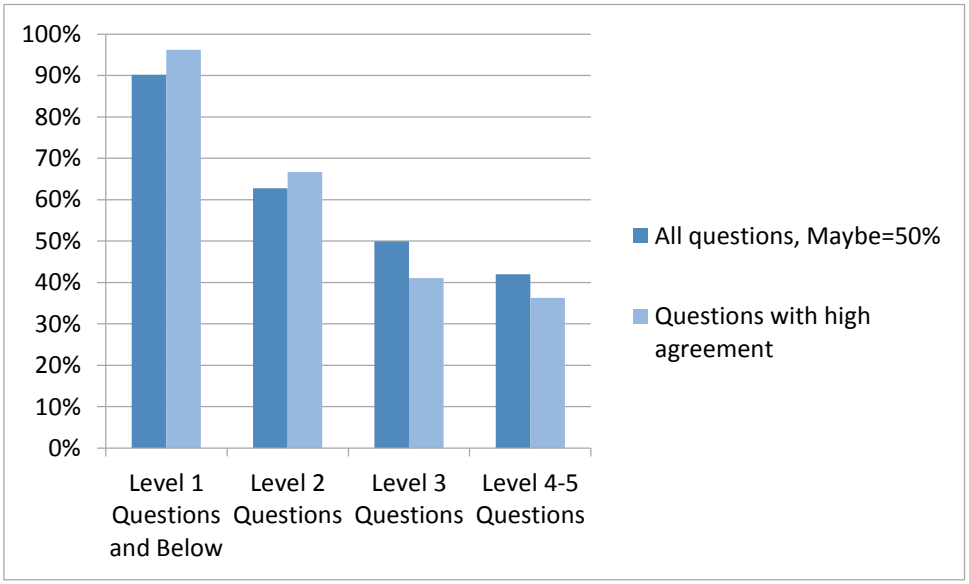
Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A4.6.

While there are differences across the different levels of question difficulty and the different ways of aggregating the ratings from the individual experts, the comparison with the aggregate human performance suggests that the literacy capabilities of current computers correspond roughly to the pattern of performance seen in Level 2 or Level 3 adults.

Disagreement on the computer literacy ratings

To examine the range of disagreement across the different questions, a simple measure of disagreement was calculated by comparing the number of Yes and No ratings. A question was identified as showing disagreement if there were at least two Yes ratings and also at least two No ratings. The Maybe ratings were ignored. Overall, 60% of the questions showed disagreement by this measure. To gauge the overall effect of disagreements on the aggregate ratings, Figure 4.7 compares the average ratings from Figure 4.1 with averages based only on the 40% of the questions where the experts showed "high agreement" - which was simply the set of questions where they did not show disagreement as defined above. The overall results using only the questions where the experts agree are quite similar to the results using all questions.

Figure 4.7 Expert ratings of computer capabilities to answer PIAAC literacy questions, comparing average using all questions to average using only questions showing high agreement, by level of question difficulty



Source: Table A4.7.

Discussion of the literacy questions

Throughout the meeting, there was extensive discussion of different literacy questions and the challenges they pose for computers. It is worth describing some of the notable exchanges that occurred to give a sense of the types of concerns and analysis that were the focus of the discussion.

The only literacy question in Level 1 and Below that the computer scientists did not agree could be answered by computers (Literacy #5) was a question with a figure that was difficult to process visually. For this question, the group divided evenly between those who believed current techniques could answer the question and those who believed they could not. The figure showed a line of people, each holding a sign showing a number. The people each represented a specific country, which was indicated by a country name positioned underneath each person, and the text indicated that the numbers on the signs represented the percentage of teachers in the country who are women. Essentially, the information in the figure could have been shown in a simple table giving the statistic for each country and in that case the computer scientists agreed that the question could have easily been answered using current computer techniques. The difficulty of the question for computers was entirely related to the problems computers would have in connecting the pieces of information in the picture.

The easiest literacy question in Level 2 (Literacy #8) raised a different kind of challenge. In this question, the test-taker sees an Internet poll related to using the Internet in cars and is given instructions to vote in the poll on behalf of another person who believes that it is unsafe for people to use Internet in their car. For this question, the group divided evenly between Yes, Maybe and No responses on the ability of computers to answer the question. The difficulty of the question in this case relates to understanding the common sense implications of the instructions - that voting on someone else's behalf means voting according to their opinion, and that voting in this context means pressing the buttons on the Internet poll website.

Another question that received extensive discussion was one of the more difficult questions in Level 3 (Literacy #44), which asks about the distance between different cities and provides a triangular distance table to use in determining the answer. Such tables are commonly used on printed maps to provide distances between pairs of cities. However, with the increasing use of computers and GPS to provide directions, many people today would never use a triangular distance table when planning a trip and some people may never have seen this kind of table format. Again the group divided evenly between Yes, Maybe and No responses on the ability of computers to answer this question. However, the discussion showed a wide range of approaches to thinking about the problem. In this case, the group did not believe that it would be hard to understand the lines and numbers of the table from the picture; instead, the issue was the ability of computers to interpret the meaning of this unusual table format. One of the computer scientists approached the question as a visual problem solving task, suggesting that the unusual format could be understood by applying standard rules for labelling more conventional tables. A number of the experts assumed that the ground rules for the test would need to specify the use of this type of table in advance, which would then make it possible to apply standard techniques during the development process to allow a computer to interpret tables of this type. One expert assumed that the information could be made available in a more standard table format and several suggested that the easiest way of answering the question would be to ignore the table provided and instead use Google to provide information for the appropriate distances.

The discussions about these three different literacy questions illustrate the wide range of factors that the computer scientists considered in determining whether current computer techniques could answer the questions. Notably, in these three questions, the difficulties that potentially prevent computers from successfully providing an answer seem to relate largely to factors other than their literacy capabilities: interpreting a difficult picture, understanding common sense implications related to voting, and having advance warning about an unusual table format. The different factors noted in these three examples are typical of a lot of the discussion that occurred around the literacy questions at the meeting, however it is possible that the non-literacy factors were discussed not because they were so important on average but because they were unusual and it was worth noting them when they occurred.

Another view of the factors being considered by the computer scientists was provided by a discussion of ten questions showing high levels of disagreement. To identify these questions, we divided the computer scientists into three groups by their overall average ratings in Figure 4.4, distinguishing the top three "Optimists" (Forbus, Burstein, and Saraswat), the bottom three "Pessimists" (Hobbs, Davis, and Graesser), and the five "Realists" in the middle (Vardi, Steedman, Passonneau, Rus, and Spohrer). The ten questions identified were those where the Optimists voted Yes (with at most one Maybe in the group) and the Pessimists voted No (with at most one Maybe in the group).[20] The Realists generally leaned towards the Optimists on the easier questions and towards the Pessimists on the harder questions.

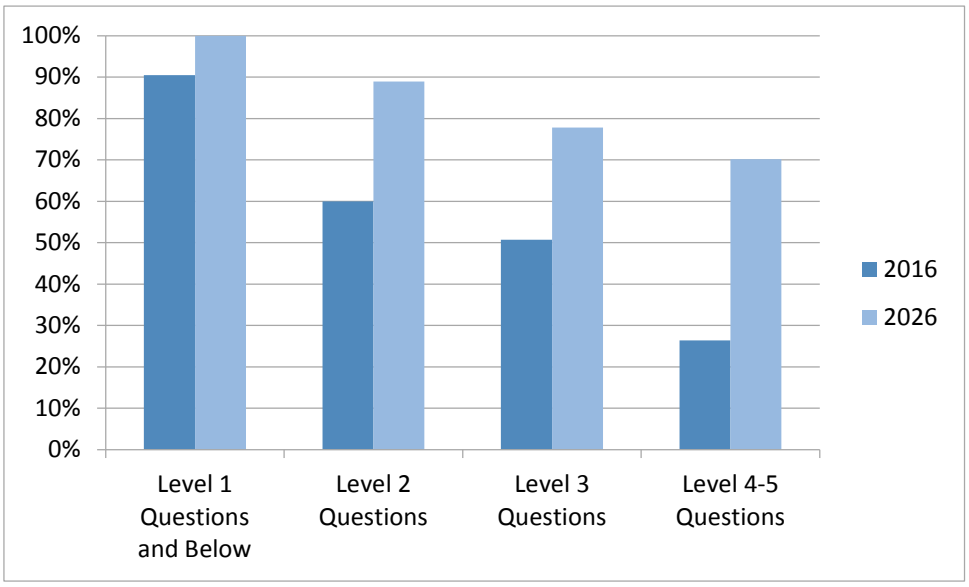
In contrast to the questions that the experts mentioned spontaneously during discussion, the comments made by the targeted discussion of the ten questions showing high disagreement between the Optimists and the Pessimists were heavily focused on issues related to language interpretation. In particular, much of the discussion concerned whether different types of "shallow" language processing would be adequate to answer each question or whether "deep" language processing would be necessary. Shallow processing involves pattern matching of various types, such as is done in search routines, while deep processing involves full interpretation of the meaning of the language. In two cases, the discussion convinced the Pessimists that the question was easier than they had originally thought and could be answered successfully with pattern matching techniques.

[20] The ten questions that meet this criterion that were identified during the meeting were 21, 23, 28, 29, 32, 39, 46, 50, 52 and 56. An additional question that meets the criterion – 35 - was identified after the meeting and so was not discussed.

Computer literacy ratings for 2026 by three experts

Three of the computer scientists also provided ratings for all of the individual questions for 2026. Although a complete analysis across all 11 experts is not possible, the partial analysis for these three provides an interesting additional perspective on expert views with respect to computer capabilities for answering the literacy test questions. The three computer scientists who provided ratings for 2026 are Davis, Forbus and Graesser, two from the Pessimist group and one from the Optimist group. The average literacy rating for 2016 for these experts is 53%, only slightly below the average rating of 56% across all 11 computer scientists. Figure 4.8 compares the average rating by proficiency level for 2016 and 2026 for these three experts, showing predicted ratings for 2026 that are substantially greater than their ratings for 2016.[21] The predicted pattern for computer performance in 2026 is somewhat better than the pattern people show who perform at Level 3 in literacy.

Figure 4.8 Comparing computer literacy ratings for 2016 and 2026, by level of PIAAC question difficulty



Source: Table A4.8.

Summary of computer ratings on the literacy questions

Overall, the computer scientists expect that computers would be more successful on the literacy questions that are easier for people and less successful on the questions that are harder for people. This pattern roughly corresponds to the increasing difficulty of the language processing required as the questions become more difficult for people, though the change in expected performance for computers across the different levels of question difficulty is weaker than it is for people. At the same time, there are notable questions at each proficiency level that the experts believe would be much more difficult for computers than they are for people. In these cases, the extra difficulty for computers often relates to additional capabilities required for the questions, such as understanding visual information or using common sense reasoning.

[21] Complete assessment ratings for computer capabilities in 2026 by literacy question and expert are provided in the Appendix in Table B4.3.

Across all 11 computer scientists, the average rating of current computer capabilities in literacy roughly corresponds to the range of performance for adults who are rated at Level 2 or 3. Such adults can answer about two-thirds of the questions at Level 2 or 3 and almost all of the easier questions. When the Maybe responses are coded as 50%, the expected pattern of aggregate performance across the different levels looks more like that of Level 2 adults. However, for the 3-expert minimum, the overall assessment of current computer capabilities looks more like the range of performance for adults who are rated at Level 3. Three computer scientists who also projected the capabilities of computers for 2026 estimated that the performance would be somewhat better than adults who perform at Level 3 in literacy.

Ratings of computer capabilities to answer the numeracy questions

Numeracy skill in the Survey of Adult Skills is defined as the "ability to access, use, interpret and communicate mathematical information and ideas, in order to engage in and manage the mathematical demands of a range of situations in adult life" (OECD, 2012). The skill includes four areas of content: quantity and number; dimension and shape; pattern, relations and change; and data and chance. The mathematical information in the test can be represented in a variety of formats, including objects and pictures; numbers and symbols; visual displays, such as diagrams, maps, graphs or tables; texts; and technology-based displays. The questions are drawn from the same familiar contexts used for the literacy test: work, personal life, society and community, and education and training.

Numeracy proficiency is described in terms of six proficiency levels, ranging from below Level 1 to Level 5. The easier test items involve single-step processes such as counting of basic arithmetic in familiar contexts. The harder test items involve complex or abstract contexts with questions requiring multiple problem solving steps related to quantitative or spatial data. For example, a Below Level 1 item has four supermarket price tags that include the packing date and asks which product was packed first. An example Level 2 item shows a logbook used by a salesman to record work-related miles of driving and asks for the reimbursement the salesman will receive for one trip noted in the logbook, using a stated reimbursement rate per mile. An example Level 4 item provides two stacked-column bar graphs showing the distribution of the Mexican population by years of schooling in different years for men and women separately and asks for one of the values shown on one of the bar graphs for one of the years and one of the categories of years of schooling. (OECD, 2013a).[22]

Computer numeracy ratings by question difficulty

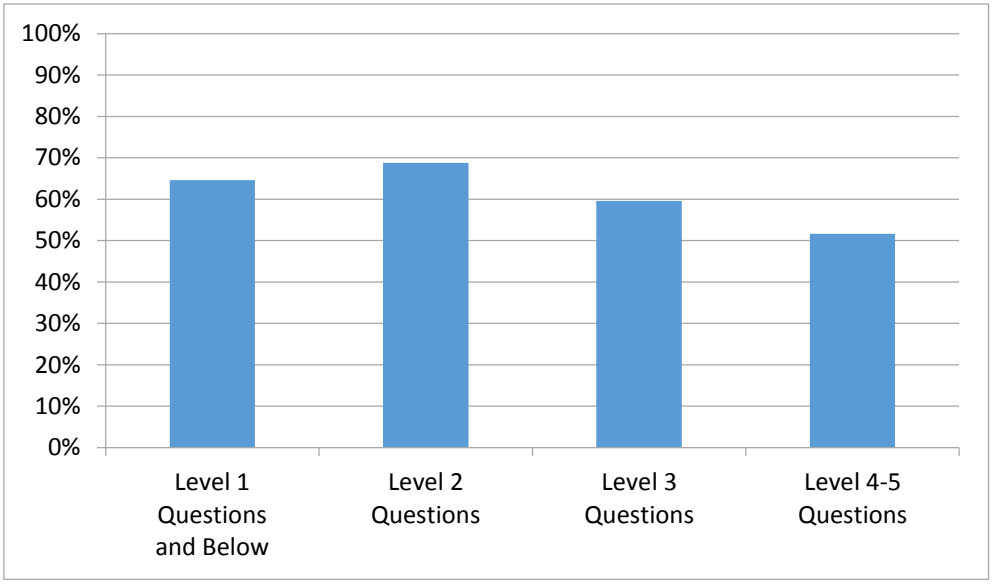
The average assessment ratings of computer capabilities on the numeracy questions are shown in Figure 4.9.[23] As with the literacy analysis, the answers of the different experts are averaged together to produce an average rating for each question and then the average ratings for all questions in each proficiency level are averaged together. The results show a relationship between the expected performance of computers and the level of difficulty of the questions for adults that is much weaker than the relationship found for literacy. For numeracy, average expected performance of current computer techniques ranges from 69% for the Level 2 questions to 52% for the Level 4 and 5 questions. Unlike the results for literacy, the expected performance of computers in numeracy for the easiest questions for adults (Level 1 and Below) is not close to 100%. Note that the particularly low rating for the questions at Level 1 and Below is almost entirely due to two questions (#1 and #8, discussed below) with difficult pictures that would be hard for a computer to interpret. The correlation coefficient across the individual questions between the

[22] More information about the Survey of Adult Skills and examples of the numeracy questions are provided in OECD (2013a, 2013b). Full descriptions of the numeracy proficiency levels are provided in the Appendix in Table B4.4.

[23] Complete assessment ratings for current computer capabilities by numeracy question and expert are provided in the Appendix in Table B4.5.

average expected rating for computers and the question difficulty score for adults is only -0.22, much smaller than the corresponding correlation for literacy.

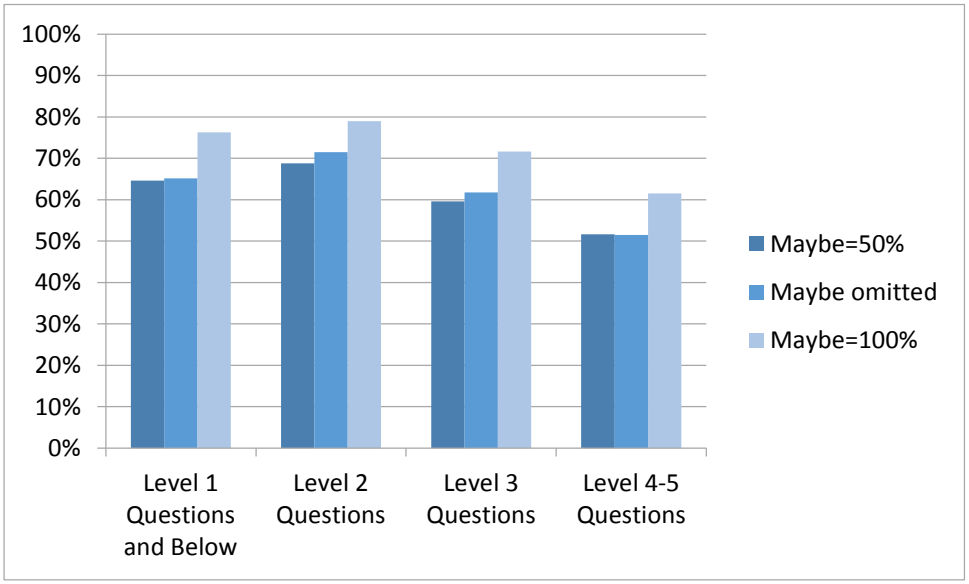
Figure 4.9 Expert ratings of computer numeracy capabilities to answer PIAAC numeracy questions, averaged with Maybe=50%, by level of question difficulty



Source: Table A4.9.

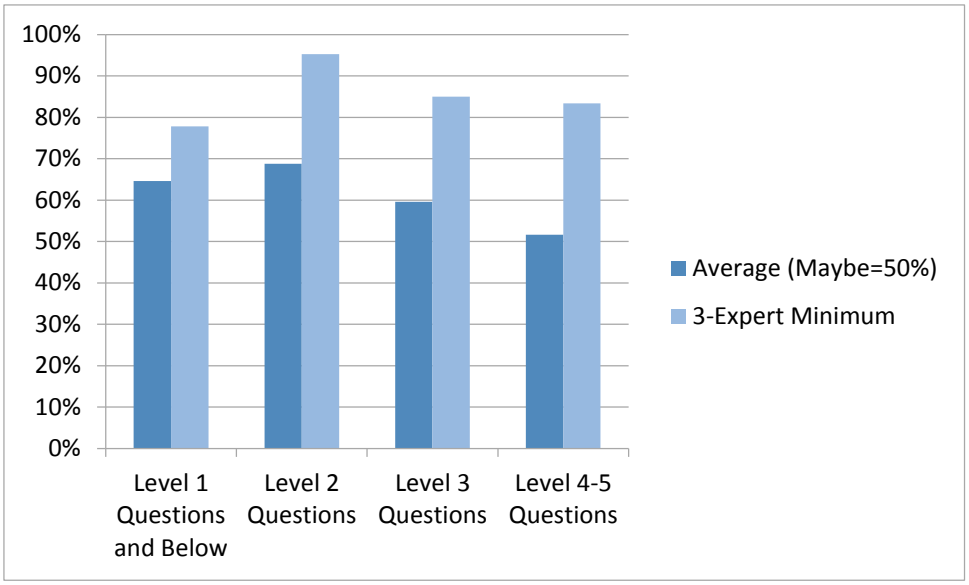
Figure 4.10 shows that alternative ratings that reflect different ways of interpreting the Maybe ratings do not produce a substantial change in the results. As with the analysis for literacy, the alternative that omits the Maybe ratings from the averages is almost indistinguishable from the version that counts Maybe as 50%. The version that counts Maybe ratings as 100% increases expected computer performance by about 10 percentage points at each numeracy proficiency level. As with the literacy, Maybe ratings accounting for a relatively small portion (22%) of the ratings.

Figure 4.10 Expert ratings of computer capabilities to answer PIAAC numeracy questions, averaged with alternative coding of Maybe ratings, by level of question difficulty



Source: Table A4.10.

Figure 4.11 Expert rating of computer capabilities to answer PIAAC numeracy questions, comparing average using Maybe=50% and 3-expert minimum, by level of question difficulty



Source: Table A4.11.

Accounting for differences in areas of expertise in the numeracy ratings

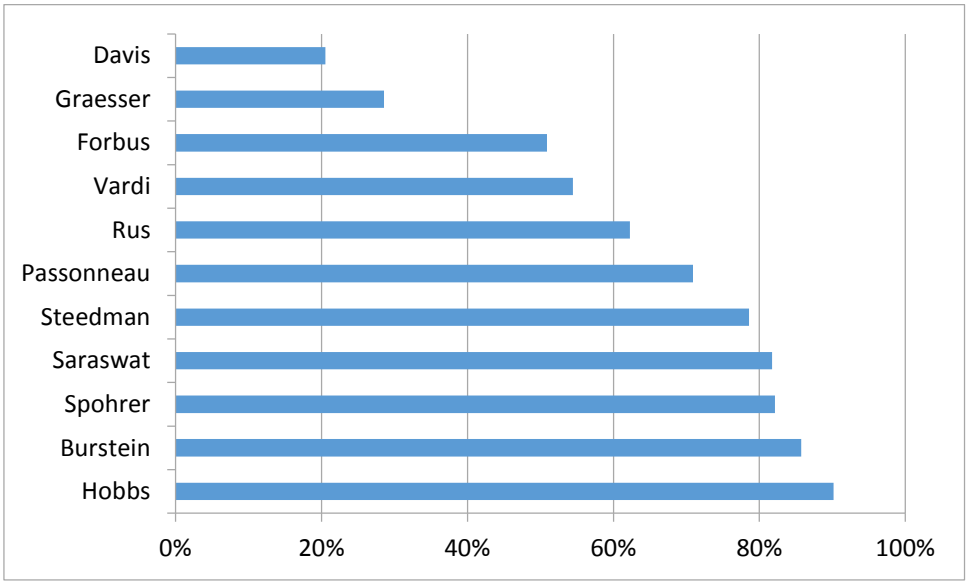
To account for differences in the areas of expertise of the experts and to allow the ratings to reflect computer capabilities from newer techniques that some experts might not yet know about, Figure 4.11 shows the results of the 3-expert minimum analysis. With this approach to aggregating the results, the proportion of questions that the group expects could be answered successfully by computers ranges

from 95% for the Level 2 questions to 83% for the Level 4 and 5 questions. As also occurred with literacy, the results from this approach suggest a substantially higher level of computer success on the numeracy questions than when the ratings are simply averaged across the group.

Computer numeracy ratings by expert

Figure 4.12 shows the average rating across all numeracy questions for each expert, counting a Yes as 100%, a Maybe as 50%, and a No as 0%. The range is 69 percentage points, from 21% for Davis to 90% for Hobbs, which is wider than the range of 56 percentage points in the ratings for literacy. The average for the group is 64%. Changing the scoring for Maybe - to omit the rating from the average or to count it as 100% - does not make an appreciable difference to the overall ratings across the different experts.

Figure 4.12 Expert ratings of computer capabilities to answer PIAAC numeracy questions, by expert



Source: Table A4.12.

Although most of the experts appear in roughly the same position in the literacy and numeracy orderings, there is a striking change for two of the experts - Hobbs and Forbus - who exchange positions: in literacy Hobbs is the most pessimistic whereas in numeracy he becomes the most optimistic; in literacy Forbus is the most optimistic whereas in numeracy he is the third most pessimistic.

As with literacy, an average for numeracy that focuses on the five experts in the middle - excluding the three most and least optimistic experts - produces results that are roughly similar to the simple average across the full group. In the case of numeracy, however, the average across the five experts in the middle tends to be somewhat higher than the average across the full group, particularly for the questions that are more difficult for people.

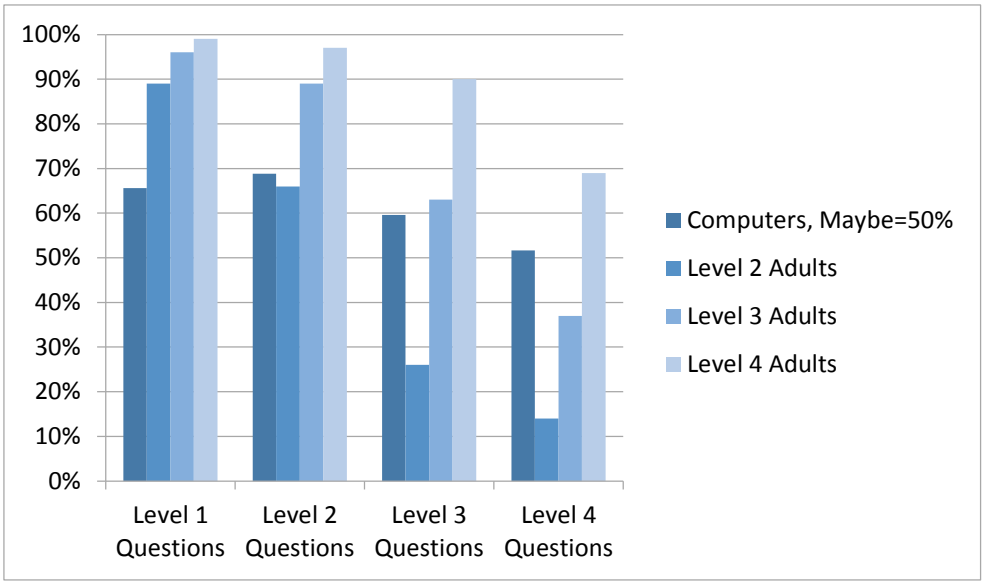
Comparing the computer numeracy ratings to human scores

Because of the lower expected performance of computers on the easiest questions in numeracy and the flatter shape of performance at the different proficiency levels, the overall pattern of expected performance looks less like the shape of typical adult performance than is the case for literacy. In general, the expected performance for computers is about 20 percentage points lower for numeracy than for literacy

on the Level 1 questions, but about 10 percentage points higher on the Level 2-4 questions (Figures 4.2 and 4.10).

Figures 4.13 and 4.14 compare the computer numeracy ratings with the performance of adults at three different levels of numeracy proficiency. Figure 4.13 uses the average ratings with Maybe coded as 50%.[24] With this coding, the computer ratings are lower than Level 2 adults for the Level 1 questions, equal to Level 2 adults for the Level 2 questions, and higher than Level 2 adults on the Level 3 and 4 questions. Figure 4.14 uses the 3-expert minimum approach that requires a minimum of three Yes ratings. With this alternative coding, the computer ratings are still lower than Level 2 adults for the Level 1 questions, but they are almost as high as the Level 4 adults for the Level 2 and Level 3 questions, and they are higher than the Level 4 adults for the Level 4 questions.

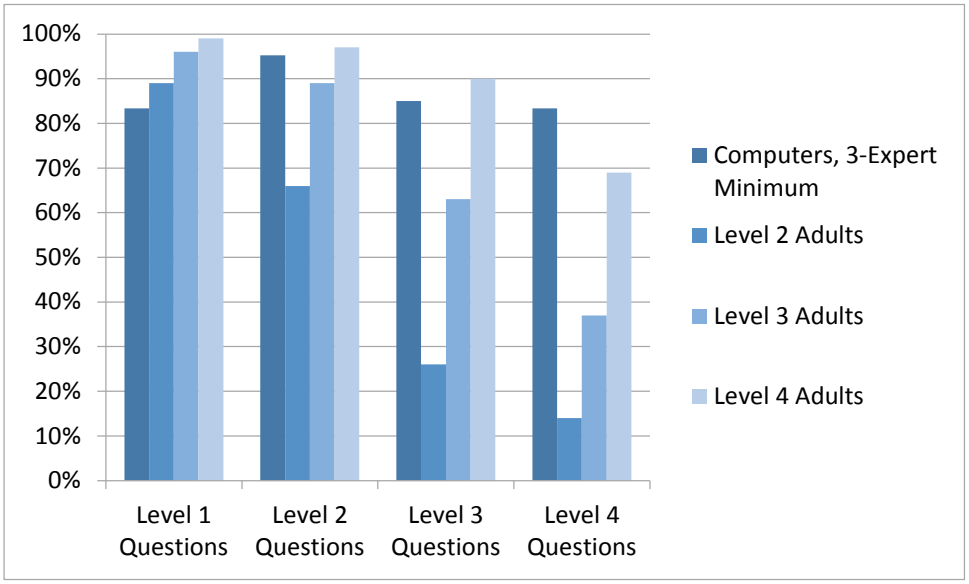
Figure 4.13 Comparing computer numeracy ratings with adults of different proficiency, using average rating with Maybe=50%, by level of PIAAC question difficulty



Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A4.13.

[24] The results are somewhat different than shown in Figure 4.9 for the computer ratings at the top and bottom because the questions below Level 1 and at Level 5 are excluded.

Figure 4.14 Comparing computer numeracy ratings with adults of different proficiency, using 3-expert minimum, by level of PIAAC question difficulty



Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A4.14.

Except for the low performance on the Level 1 questions, the comparison with human performance suggests that the numeracy capabilities of current computers correspond roughly to the pattern of performance seen in Level 2 or 4 adults, depending on the method used to aggregate the individual responses from the experts.

Disagreement on the computer numeracy ratings

The experts showed a somewhat higher level of disagreement on the numeracy questions than the literacy questions. Overall 66% of the questions showed disagreement for numeracy, compared to 60% for literacy, using the same measure where a question is identified as showing disagreement if there are at least two Yes ratings and also at least two No ratings. Most of the questions (88%) in Levels 3-5 showed disagreement. Because of the small number of questions where the experts agree in Level 3 and Levels 4-5, it is not meaningful to compare the results by numeracy proficiency level using only the questions where the experts agree.

Discussion of the numeracy questions

During the discussion of numeracy, a number of the experts noted that successfully applying computer techniques to answer the numeracy questions seemed to require the development of a large number of specialized systems to address particular types of questions and to process particular types of figures, tables or pictures. Although in most cases the development of any one of these specialized systems did not seem to be a problem, it was unclear how many specialized systems would need to be created to answer the full set of possible questions on the test. Without a well-defined specification of the types of material that might be presented, the example test questions suggested that the number of potential types of questions that might be used in the test could be quite large. This requirement for developing a number of specialized systems for numeracy contrasted with the situation for literacy, where the experts believed that many of the questions could be addressed with a relatively small number of general language techniques.

Much more of the discussion of individual numeracy questions ended up focusing on issues related to understanding the visual input for the different types of questions. As a result, several of the experts mentioned feeling less confident about their judgments about the numeracy questions because they felt they did not have sufficient expertise to evaluate the visual processing requirements for the different questions.

The numeracy question that is the easiest for people (Numeracy #1) was mentioned repeatedly during the discussion because of the striking contrast between the expected low performance for computers and the high performance for people. As noted in Chapter 3, this question received the lowest rating for computer capability across the group and it was the only numeracy question that did not receive any Yes votes. The experts thought that the difficulty of this problem related primarily to the difficulty of interpreting a photograph of two packages of bottled water, because the packaging material makes it hard to identify many of the bottles in the photograph. The difficulty of the mathematical reasoning required to determine how many bottles are in the packages was not the feature that would make the question hard for computers.

Another numeracy question in Level 1 (Numeracy #8) received very low ratings for similar reasons. This question uses a photograph of a box of candles and asks how many layers of candles are in the box. As with the photograph of the packaged water bottles, the photograph of the packaged candles is hard to interpret because many of the candles are not directly visible and must be inferred. So the difficulty of the question relates to the difficulty of interpreting the photograph, not the difficulty of the mathematical reasoning involved to answer the question about how many layers of candles are in the box.

As with the literacy discussion, the group discussed a set of numeracy questions that showed disagreements between the three top Optimists (Hobbs, Burstein, Spohrer) and the three top Pessimists (Davis, Graesser, Forbus) in the group. The group discussed 8 of the 16 questions identified where the Optimists voted Yes (with at most one Maybe in the group) and the Pessimists voted No (with at most one Maybe in the group).[25] The five experts in the middle leaned towards the Optimists on most of these questions. Half of the questions discussed raised issues related to visual materials that the experts believed would be difficult for computers to interpret. In addition, another one of the questions asked the test-taker to use a ruler to measure a line and the group decided that they lacked the necessary expertise in robotics to evaluate the relevant computer capabilities. Unlike the corresponding discussion about the literacy questions, there were no cases in numeracy where the Optimists and Pessimists ended up agreeing on their rating for a question after discussing it.

Computer numeracy ratings for 2026 by three experts

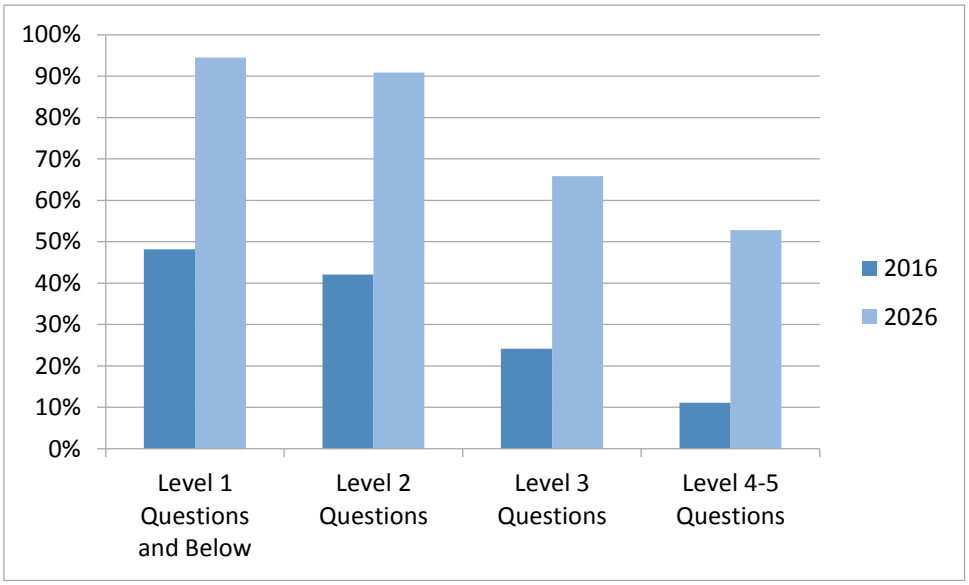
Three of the computer scientists also provided ratings for the numeracy questions for 2026. As for literacy, the three computer scientists who provided ratings for 2026 are Davis, Forbus and Graesser. Since Forbus moved to the Pessimist group for numeracy, all three of the experts who provided ratings for 2026 were in the Pessimist group. The average literacy rating for 2016 for these experts is 33%, substantially below the average rating of 64% across all 11 computer scientists. Figure 4.15 compares their average rating by numeracy proficiency level for 2016 and 2026, showing a substantial expected increase in computer capabilities over the ten-year period.[26] The projected increase is much larger for numeracy than it was for literacy. With respect to human skills, the predicted pattern of performance

[25] The 16 questions that meet this criterion that were identified during the meeting were 5, 16, 17, 25, 28, 31, 34, 36, 37, 41, 42, 43, 49, 51, 52 and 54. During the discussion of disagreements, the group discussed questions 16, 17, 25, 28, 31, 34, 36 and 41.

[26] Complete assessment ratings for computer capabilities in 2026 by numeracy question and expert are provided in the Appendix in Table B4.6.

in 2026 is close to the pattern that people show who perform at Level 3 in numeracy proficiency, expecting success on about two-thirds of the questions at Level 3 and almost all of the easier questions at Level 2 and below.

Figure 4.15 Comparing computer numeracy ratings for 2016 and 2026, by level of PIAAC question difficulty



Source: Table A4.15.

Summary of computer ratings on the numeracy questions

Unlike the ratings for the literacy questions, the computer scientists expect that computer performance will show only a small difference between the numeracy questions that are easier for people and those that are more difficult. In general, the nature of the mathematical reasoning required for different questions was seldom raised as a difficulty in the group's discussion. Instead, the primary focus was on the difficulties presented by the different visual materials and by particular problem types. In a few cases, the experts also mentioned challenges related to the use of language in understanding the question or the text.

The average rating of current computer capabilities in numeracy is somewhat difficult to compare to the performance for adults because the predicted computer performance is relatively similar at the different levels. Although the group projected that current computers could be successful on about two-thirds of the numeracy questions at Levels 2, 3 or even 4, depending on the aggregation method used, they did not expect that computers could be successful on most of the easiest questions at Level 1 and Below. The primary problem posed for computers by the easiest questions for adults was the interpretation of visual material. Finally, three computer scientists who projected the capabilities of computers for 2026 estimated that the performance on numeracy would be similar to adults who are rated at Level 3; these three experts were the ones who gave the lowest overall ratings in the group for the computer performance on numeracy in 2016.

Ratings of computer capabilities to answer the problem solving questions

Skill in problem solving with computers[27] in PIAAC is defined as "using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks" (OECD, 2012). The domain involves the ability to solve problems for personal, work or civic purposes by setting up goals and plans, and accessing and making use of information through computers. Although the skill area is intended to address a full range of digital devices, the current version of the test is limited to work on a laptop computer using generic versions of email, browser and spreadsheet software.

Problem solving proficiency is described in terms of four proficiency levels, ranging from below Level 1 to Level 3. The easier test items involve well-defined problems using only a single function of one of the generic programs and without any inference required. The harder test items involve combining multiple steps across multiple programmes to solve a problem where the goal may not be fully defined and unexpected outcomes may occur. For example, a Level 1 item asks the test-taker to sort email responses to a party invitation into two existing folders for those who can and cannot attend. An example Level 2 item asks the test-taker to respond to an email asking about club members who meet two conditions using a spreadsheet containing 200 entries describing each of the members. An example Level 3 item involves multiple email requests to reserve meeting rooms using a web-based reservation system and resolving a conflict related to two of the requests (OECD, 2013a).[28]

The six computer scientists who provided ratings for the problem solving domain are Davis, Forbus, Graesser, Passonneau, Spohrer and Steedman. In literacy, these six experts gave an average rating of 56%, the same as the average for all 11 experts. In numeracy, these six experts gave an average rating of 55%, somewhat below the average of 64% for all 11 experts. The results for the other two skill areas suggest that these six experts are likely to give a set of average ratings for the problem solving domain that are roughly comparable to the average that would have resulted from the full group of 11 computer scientists.

Computer problem solving ratings by question difficulty and by expert

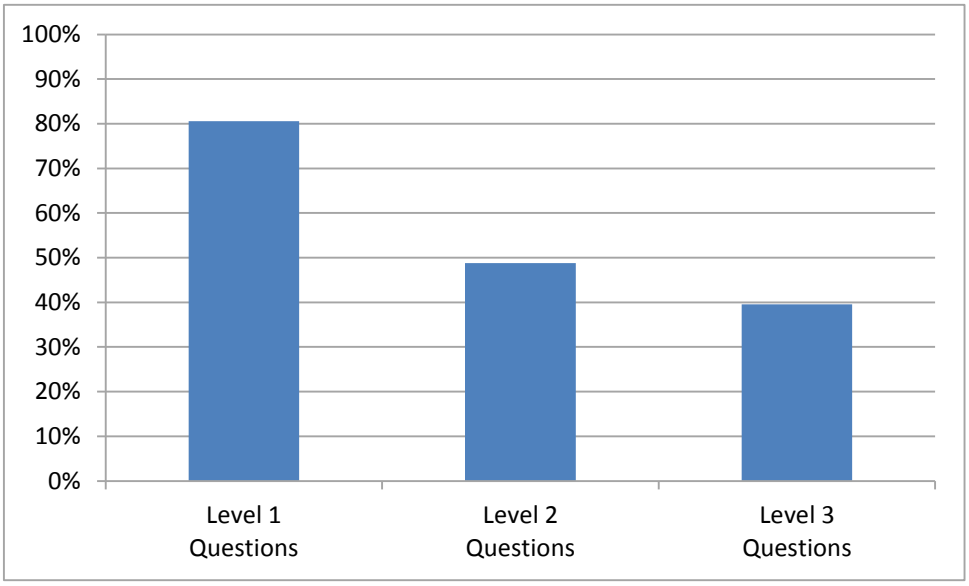
Figure 4.16 provides the average expert ratings of computer capabilities to answer the questions in problem solving with computers for each proficiency level.[29] As for the other skill areas, the answers of the different experts are averaged together to produce an expected result for each question and then the average expert ratings for all the questions in each proficiency level are averaged. The results show a relatively strong relationship between the expected performance of computers and the level of difficulty of the questions for adults in this domain. The correlation coefficient across the individual questions between the average expected rating for computers and the question difficulty score for adults is -0.74. The results in Figure 4.16 average the individual ratings coding Maybe as 50%. The versions with Maybe omitted, with Maybe coded as 100%, or with requiring a minimum of three Yes ratings all produce similar results.

[27] The formal term used for this domain in PIAAC is "problem solving in technology-rich environments."

[28] More information about the Survey of Adult Skills and examples of the problem solving questions are provided in OECD (2013a, 2013b). Full descriptions of the problem solving proficiency levels are provided in the Appendix in Table B4.7.

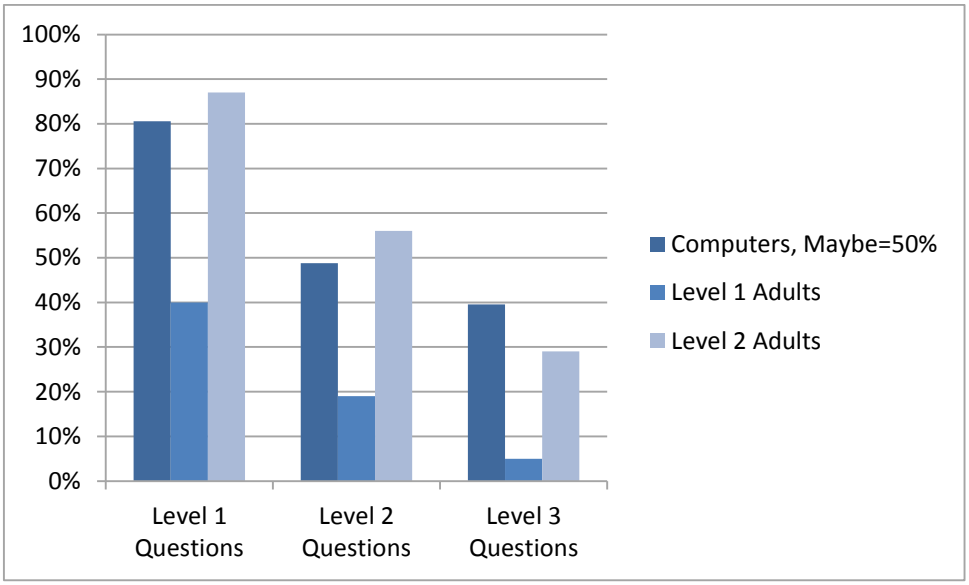
[29] Complete assessment ratings for current computer capabilities by problem solving question and expert are provided in the Appendix in Table B4.8.

Figure 4.16 Expert ratings of computer capabilities to answer PIAAC problem solving questions, averaged with Maybe=50%, by level of question difficulty



Source: Table A4.16.

Figure 4.17 Comparing computer problem solving ratings with adults of different proficiency, using average rating with Maybe=50%, by level of PIAAC question difficulty



Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A4.17.

Figure 4.17 compares the expected computer ratings with the performance of adults at two different levels of proficiency in problem solving using computers.[30] The shape of the experts' expectations of

[30] The problem solving skill area is scored using only three levels of difficulty because of the small number of test questions, rather than the five levels used in literacy and numeracy.

computer capabilities across the different proficiency levels relatively closely matches adults with a proficiency of Level 2 in problem solving with computers.

The ratings of the six experts across all problem solving questions range 93 percentage points, from 0% for Graesser to 93% for Passonneau. The average rating across all six experts and all questions is 53%, substantially lower than numeracy and slightly lower than literacy. The range of disagreement for the problem solving domain is wider than either of the other two domains. However, given the smaller number of experts, further analyses about the level of disagreement were not conducted.

Discussion of the problem solving questions

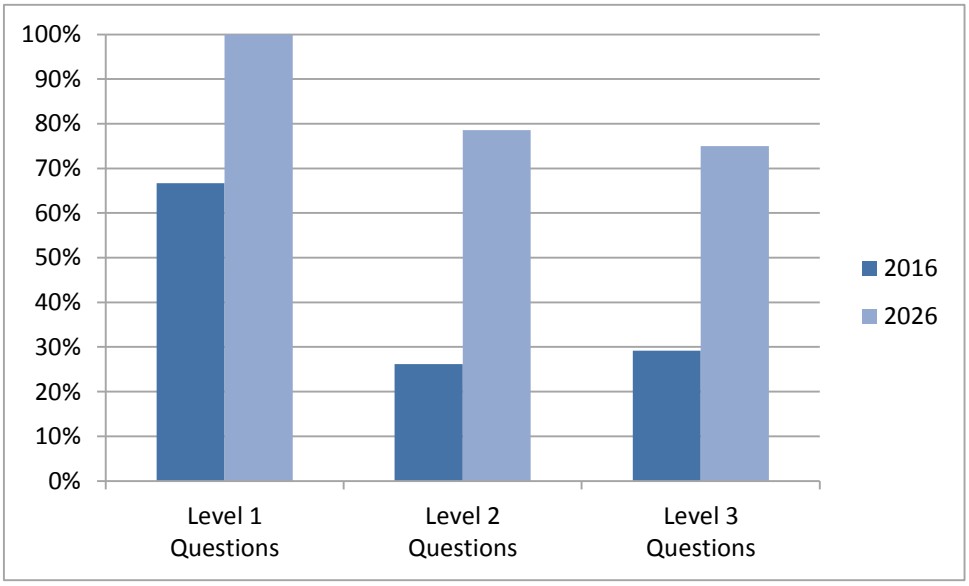
The group did not have time at the meeting to discuss the questions in the domain of problem solving with computers. However, notes prepared by the participants in advance contain several points related to this domain. Some of the experts expected that the context of the different questions would be difficult to interpret and that this would cause the problem solving questions to be more difficult for computers than the literacy and numeracy questions, though this did not turn out to be the case in the actual ratings. Many of the specific points raised in the advance notes related to issues of language understanding rather than expected difficulties related to problem solving or the use of software applications.

Computer problem solving ratings for 2026 by three experts

As for the other two domains, three of the computer scientists also provided ratings for 2026 for the questions for problem solving with computers. Davis and Forbus were in the middle of the expert distribution for 2016, whereas Graesser was at the bottom. Overall, these three experts had an average score for 2016 of 36%, below the average rating of 53% across all six computer scientists who provided ratings for problem solving. Figure 4.18 compares the average rating by the proficiency level for problem solving for 2016 and 2026 for the three experts who provided both.[31] As with the ratings for literacy and numeracy, the predicted capability ratings for 2026 for problem solving are substantially greater than the corresponding ratings for 2016. The predicted pattern of performance is better than the pattern that people show who perform at Level 2 in problem solving with computers, and is almost as good as Level 3, which is the highest performance level on the test.

[31] Complete assessment ratings for computer capabilities in 2026 by problem solving question and expert are provided in the Appendix in Table B4.9.

Figure 4.18 Comparing computer problem solving ratings for 2016 and 2026, by level of PIAAC question difficulty



Source: Table A4.18.

Summary of computer ratings on the problem solving questions

Although only half of the experts provided ratings for the problem solving questions and the group did not have a chance to discuss the domain at the meeting, the available ratings provide an initial sense of the capabilities of computers in this area. Like the ratings for literacy and unlike those for numeracy, the computer scientists expect that computer performance will be stronger on the questions that are easy for people and weaker on the questions that are harder for people. Across the six computer scientists who provided ratings, the average rating of current computer capabilities in problem solving with computers roughly corresponds to the range of performance for adults who are rated at Level 2 in this skill area. Three computer scientists who also projected the capabilities of computers for 2026 estimated that the performance in the problem solving domain at that time would be almost as good as the top adult performance level on the test.

5 – IMPLICATIONS OF COMPUTER CAPABILITIES FOR POLICY AND RESEARCH

The preceding chapters of this report have presented two substantially different analyses: first, a discussion of past changes in literacy skills and skill use in Chapter 2, and then a discussion of current computer capabilities in literacy and other general cognitive skills in Chapters 3 and 4. This chapter joins those two analyses to consider how computer capabilities in general cognitive skill areas are likely to change the use of those skills in the workplace in the future. This consideration has implications for the development of general cognitive skills, as well as the ways that skills are assessed for shaping education and labour policy.

Linking current computer capabilities to workforce skill trends

The exploratory assessment of computer capabilities described in Chapters 3 and 4 resulted in several different aggregate ratings for each of the three skill areas included in the Survey of Adult Skills (PIAAC). The discussion in Chapter 4 compared these different ratings for computers to adults at different proficiency levels, which involved looking at the difference between expected computer performance and actual adult performance for questions at different levels of difficulty. In general, human performance decreases more steeply than computer performance does as the questions become more difficult for people, so some approximation is required to choose a human proficiency level that roughly corresponds to expected computer performance. Table 5.1 summarizes the proficiency levels identified in Chapter 4 that correspond to the three different skill areas and three of the aggregate computer ratings. These ratings should be treated as preliminary, resulting from an exploratory process that did not have sufficient time to try several proposed ways of resolving the disagreements between the experts about their judgments. However, it is worth taking this set of aggregate ratings at face value and considering their implications for workplace skills.

Table 5.1 Approximate proficiency level of computer capabilities on PIAAC

Computer Rating	Literacy	Numeracy	Problem Solving
Current capabilities, average with Maybe as 50%	Level 2	Level 2	Level 2
Current capabilities, 3-expert minimum	Level 3	Level 4	
Capabilities in 2026	Level 3	Level 3	Level 3

Source: Tables A4.5, A4.6 A4.8, A4.13, A4.14, A4.15, A4.17, A4.18.

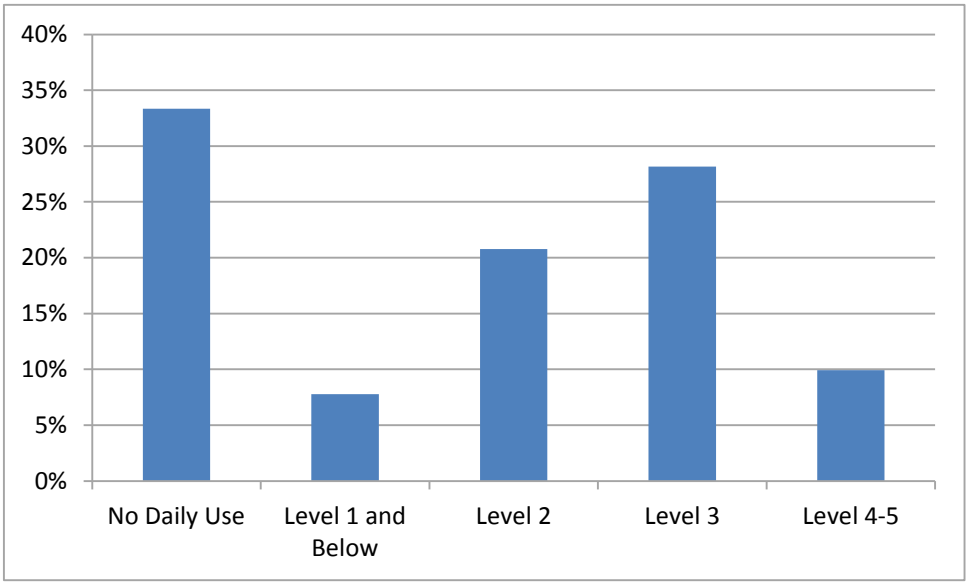
The broadest aggregate rating in Table 5.1 is the simple average that counts Maybe as 50%, which reflects the judgments of the full set of computer scientists for literacy and numeracy.[32] This group included experts with a number of different specialties and a range of overall optimism about computer capabilities. When this rating has a high value, it means that most of the group was able to suggest current

[32] Only three computer scientists provided ratings for the third skill area of problem solving with computers.

approaches that they believed would allow computers to answer a particular question. The other two aggregate ratings in the table reflect a smaller group of experts. The 3-expert minimum rating generally reflects more optimistic experts, since it requires only three of the experts to indicate that a question could be answered by computers. And the 2026 rating was made by only three members of the group, who turned out to be somewhat more pessimistic than the group as a whole. So the first aggregate rating provides a relatively conservative judgment that required agreement from a broad set of experts that a question could be answered by computers with today's capabilities. In contrast, the second and third ratings provide two different ways of thinking about the boundaries of what may be possible in the near future, either because several more optimistic experts say these things can be done today or because several more pessimistic experts say these things could be done ten years from now.

For literacy, the first computer rating is for Level 2, and the second and third are both for Level 3. For comparison, Figure 5.1 shows the proportion of workers at different levels of literacy proficiency who use literacy on a daily basis, averaged over all OECD countries and economies included in the Survey of Adult Skills. This figure is similar to the PIAAC results in Figure 2.8 from Chapter 2, although the average includes data for ten countries that are not included in the earlier figure because they did not participate in IALS.[33] The first rating of Level 2 suggests that the literacy-related tasks of 29% of the workforce could be affected by current computer capabilities, whereas the second and third ratings suggest that the literacy-related tasks of 57% of the workforce could be affected. In addition, 43% of the workforce would not be strongly affected by these computer capabilities, either because literacy-related materials are not a daily part of their work or because their literacy proficiency is above the level that computers will be able to provide in the near future.

Figure 5.1 Distribution of workers by daily literacy use and level of proficiency



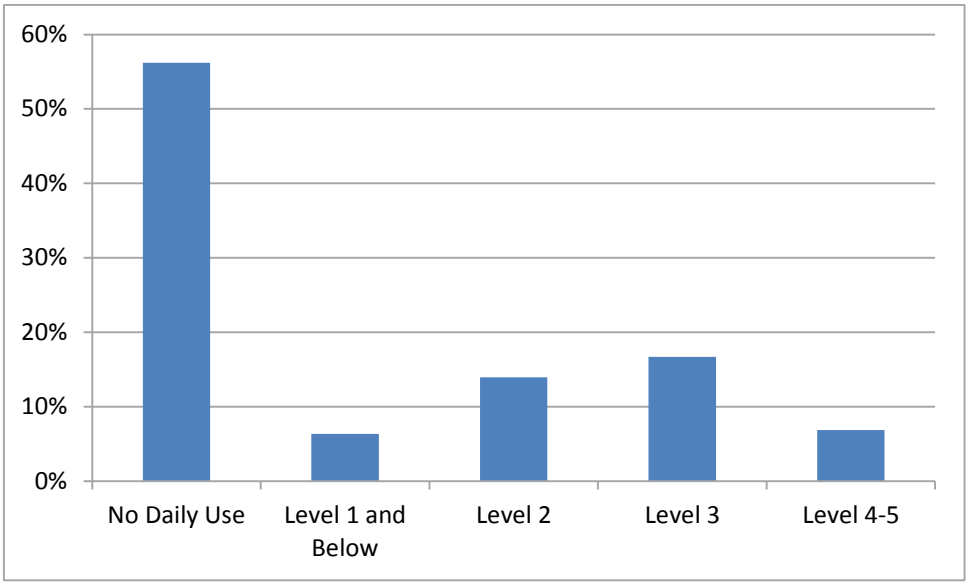
Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A5.1.

For numeracy, the first computer rating is for Level 2, and the second and third are for Levels 4 and 3, respectively. This provides a wider range of worker proficiency levels that could potentially be affected by computer capabilities. Figure 5.2 shows the proportion of workers at different levels of numeracy

[33] The additional countries are Austria, Estonia, France, Greece, Israel, Japan, Korea, the Slovak Republic, Spain, and Turkey.

proficiency who use numeracy on a daily basis.[34] Not surprisingly, there are fewer workers who use numeracy than who use literacy on a daily basis. The numeracy-related tasks of 20% of the workforce could be affected by current computer capabilities of Level 2. This increases to 37% for computer capabilities at Level 3 and to 44% at Level 4, effectively the entire workforce that uses numeracy on a daily basis at work.[35]

Figure 5.2 Distribution of workers by daily numeracy use and level of proficiency



Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A5.2.

For problem solving with computers, the first computer rating is for Level 2 and the third is for Level 3.[36] Figure 5.3 shows the proportion of workers at different levels of proficiency in problem solving with computers who use computers on a daily basis.[37] More than three-quarters of workers use computers on a daily basis at work. The computer-related tasks of 69% of the workforce could be affected by current computer capabilities of Level 2.[38] This increases to all workers using computers on a daily basis (76%) with computer capabilities of Level 3.

[34] To identify daily numeracy use, the analysis aggregates skill use questions related to reading "bills, invoices, bank statements or other financial statements," reading "diagrams, maps or schematics", calculating "prices, costs or budgets," and using or calculating "fractions, decimals or percentages."

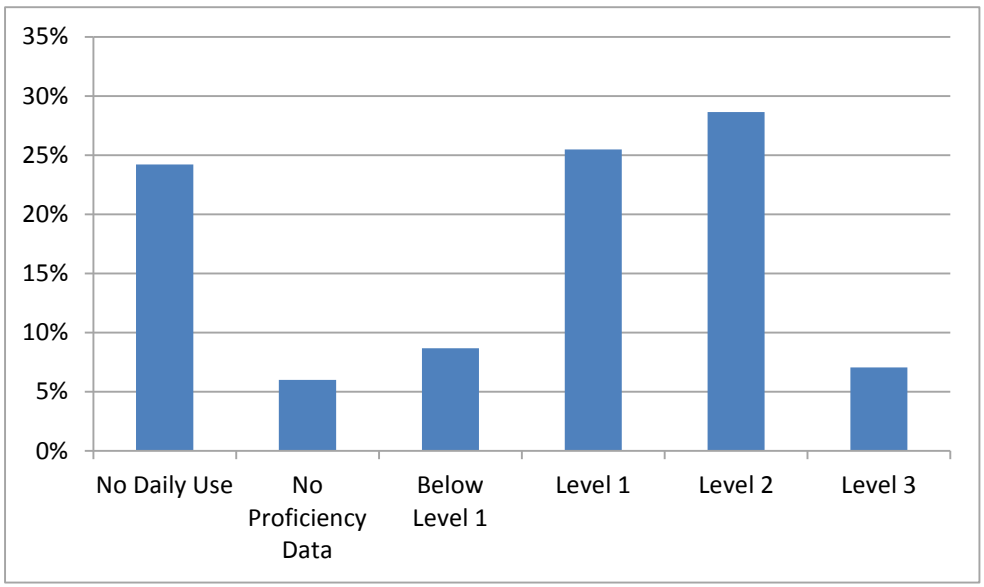
[35] Level 5 represents only 1% of the population in the OECD average (OECD, 2016b), so the question about whether or not computer capabilities would be able to reach performance in numeracy comparable to Level 5 would not have a substantial effect on the portion of the workforce affected.

[36] No 3-expert minimum analysis was performed for the problem solving dimension because only three experts provided responses.

[37] To identify daily computer use, the analysis aggregates skill use questions related to using "email," using "the internet in order to better understands issues related to your work," conducting "transactions on the internet, for example buying or selling products or services, or banking," using "spreadsheet software," and using "a word processor."

[38] The category of workers who use computers on a daily basis but have no proficiency data includes those who failed the initial screening test related to basic computer operation or who opted out of the computer test.

Figure 5.3 Distribution of workers by daily computer use and level of proficiency

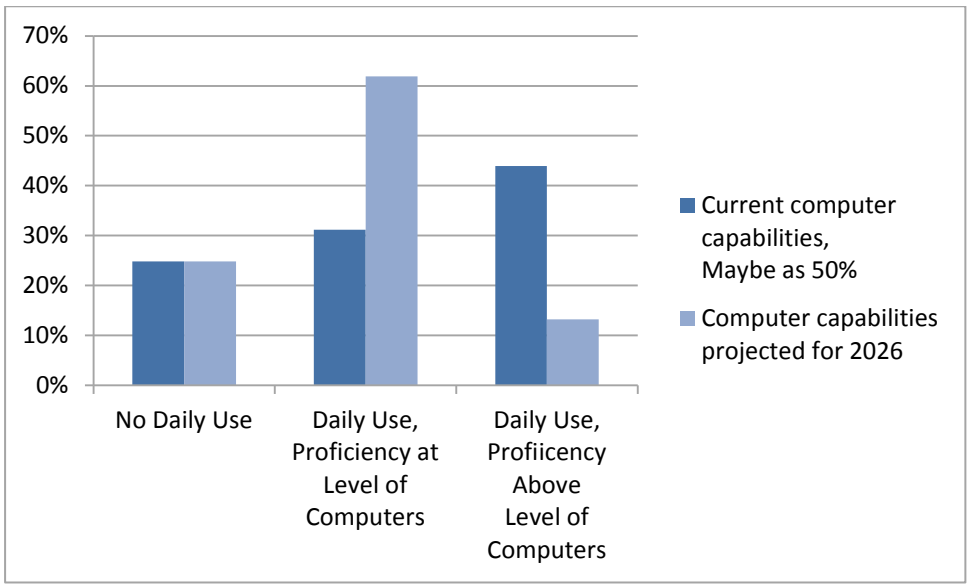


Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A5.3.

Finally, Figure 5.4 combines the analyses for these three general cognitive skills to identify the portion of the workforce that will be affected by these computer capabilities for the first and third computer ratings. At the lower end, 25% of the workforce does not use any of these three skills on a daily basis at work, so their regular work tasks will not be substantially affected by these particular computer capabilities. At the upper end, there are workers who use one or more of these skills on a daily basis and have proficiency above the projected level of computer capabilities. Because these workers have proficiency in the three skill areas that are above projected computer capabilities for the near future, it is reasonable to expect they will continue to have regular work tasks using these skills that are not substantially affected by these particular computer capabilities. This proportion is 44% for the projected level of computer capabilities in 2016 using the first rating, and 13% for the projected level of computer capabilities in 2026 using the third rating. In the middle, between the workers who do not regularly use these skills and the workers who regularly use them at a level above the projected capabilities of computers, there is a large proportion of workers who use one or more of these skills on a daily basis but have proficiencies only at the level of projected computer capabilities. This proportion is 31% for the projected level of computer capabilities in 2016 using the first rating, and 62% for the projected level of computer capabilities in 2026 using the third rating.

The calculation assumes that essentially none of these computers would be at the highest level of proficiency on the assessment of problem solving with computers if they had attempted it.

Figure 5.4 Distribution of workers by use of general cognitive skills and proficiency compared to computers

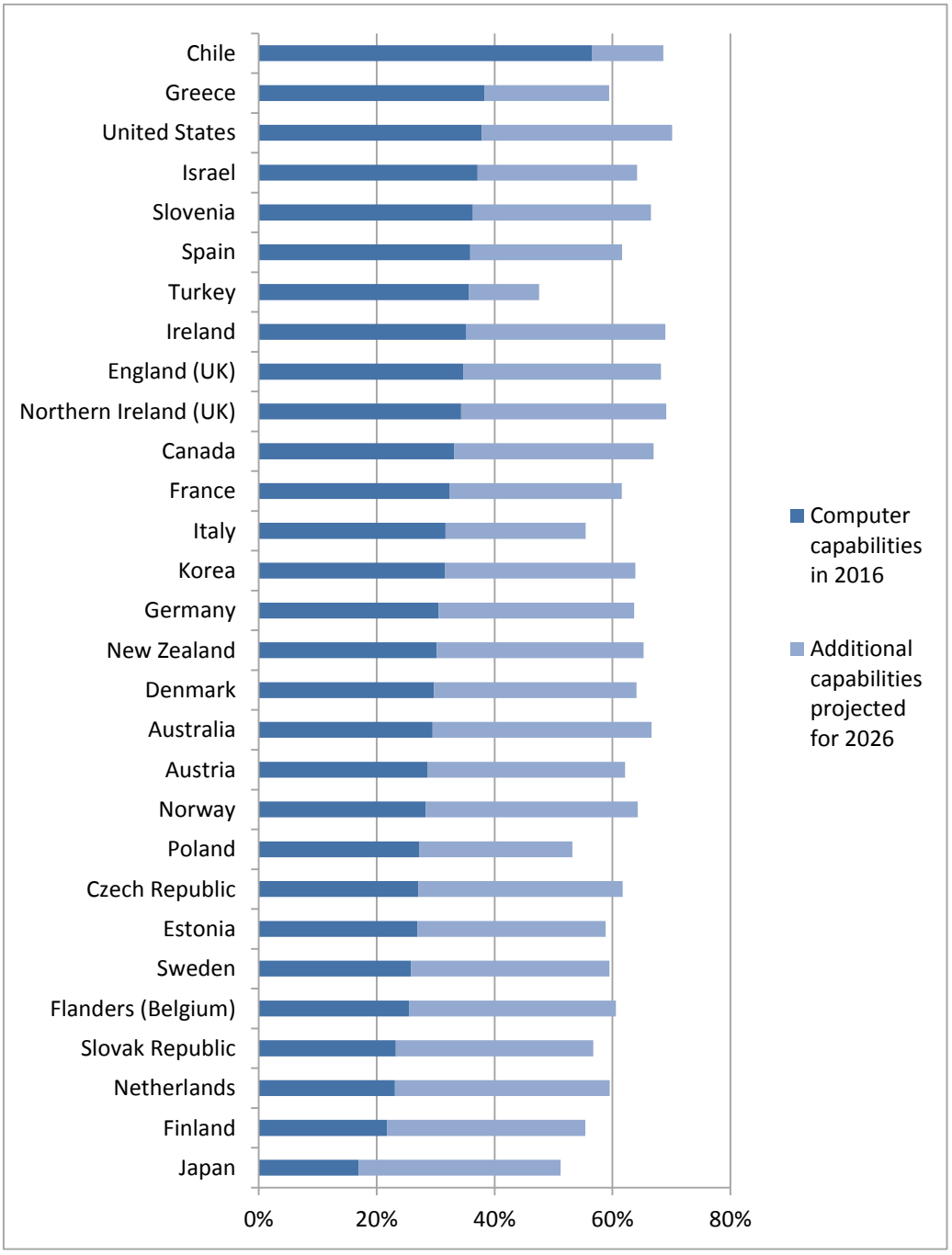


Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A5.4.

The workers in the middle are the ones whose work tasks seem most likely to be substantially affected by the projected computer capabilities in these three areas of general cognitive skill. Figure 5.4 suggests that the next two decades are likely to see a reversal in the pattern of skill use change that Chapter 2 described for the last two decades, at least with respect to the three general cognitive skills measured by PIAAC. Between IALS and PIAAC, the proportion of the workforce using written materials on a daily basis increased for those with a low to medium level of proficiency, while decreasing for those who do not use written materials regularly and for those with a high level of proficiency. However, in the next two decades, computers will be able to substitute for people with a low to medium level of proficiency in carrying out many of the tasks using these general cognitive skills. Although the changes in literacy proficiency and use over the past two decades increased the proportion of the workers using literacy regularly and having only low or mid-level skills, the assessment of computer capabilities in the three general skill areas measured by PIAAC suggest that workers with only low to mid-level proficiency in these skills may be less likely to use them regularly at work in the coming decades.

There are large differences across countries in the proportion of the workforce that regularly uses these three skills with proficiency levels at or below the capabilities projected for computers. Figure 5.5 shows that the potentially affected workforce ranges from 17% for Japan to 56% for Chile using the first rating for 2016. With the third rating for 2026, the potentially affected workforce ranges from 48% for Turkey to 70% for the United States. Some countries, like Chile, show many workers potentially affected by computers using this measure because they have relatively few workers with proficiency above the projected level of computer capabilities; other countries, like the United States, show a high proportion potentially affected because more workers regularly use these skills at work.

Figure 5.5 Proportion of workforce using general cognitive skills with proficiency at or below level of computer capabilities



Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A5.5.

Implications of computer capabilities for employment

The preceding comparison of the computer capability projections with the proficiency and skill use of the work force raises questions about how these computer capabilities will affect employment. Although there is not sufficient information for a full answer, several points can be made.

First, the analysis is only preliminary. Chapters 3 and 4 identify a number of limits that affected the exploratory judgments of computer capabilities that form the basis for the projection of affected workers.

Second, the analysis focuses on technical capability rather than economic application. It is well established that the application of new technologies often takes a decade or more when it occurs - and sometimes it never occurs (Comin and Hobijn, 2010; Griliches, 1957; Mansfield, 1961). A further analysis of the economic and organizational factors that will affect the application of the projected computer capabilities would need to be carried out to understand which applications are likely to take place and which are not. The existing literature also suggests that the speed of diffusion is likely to vary substantially by country (Comin and Hobijn, 2010) and by firm (OECD, 2015a).

Third, most jobs involve a mix of different types of skill and can vary considerably in the relative importance of the different skills and how closely they are linked together. This has implications for the technical scope for using computers that have capabilities in some skill areas but not others to automate job tasks. For example, most receptionists, nurses and housekeepers regularly use both language and physical skills, but the role of those skills is different in each job: many receptionist tasks could be automated with language skills alone, whereas many housekeeping tasks could be automated with physical skills alone, and many nursing tasks require both language and physical skills. Without knowing computer capabilities in other skill areas and the skill mix required in different jobs, it is hard to know how computer capabilities in the three general cognitive skills alone would affect employment.

Despite the difficulties in drawing clear conclusions about employment effects, the level of current computer capabilities does suggest that further increases in the use of low and medium general cognitive skills are unlikely in the next several decades. Even if there are not significant decreases in the employment of workers with skills at these levels, it would be prudent to expect that the demand for general cognitive skills at these low and medium levels will become weaker. It is quite likely that many of workers with general cognitive skills at these levels will still be employed, but they will be employed primarily because of other skills that they have - perhaps physical skills, social skills, or special expertise in some particular content area. This has implications in turn for how researchers and policymakers should analyse and understand skill development.

Realistic aspirations for general cognitive skill development in the general population

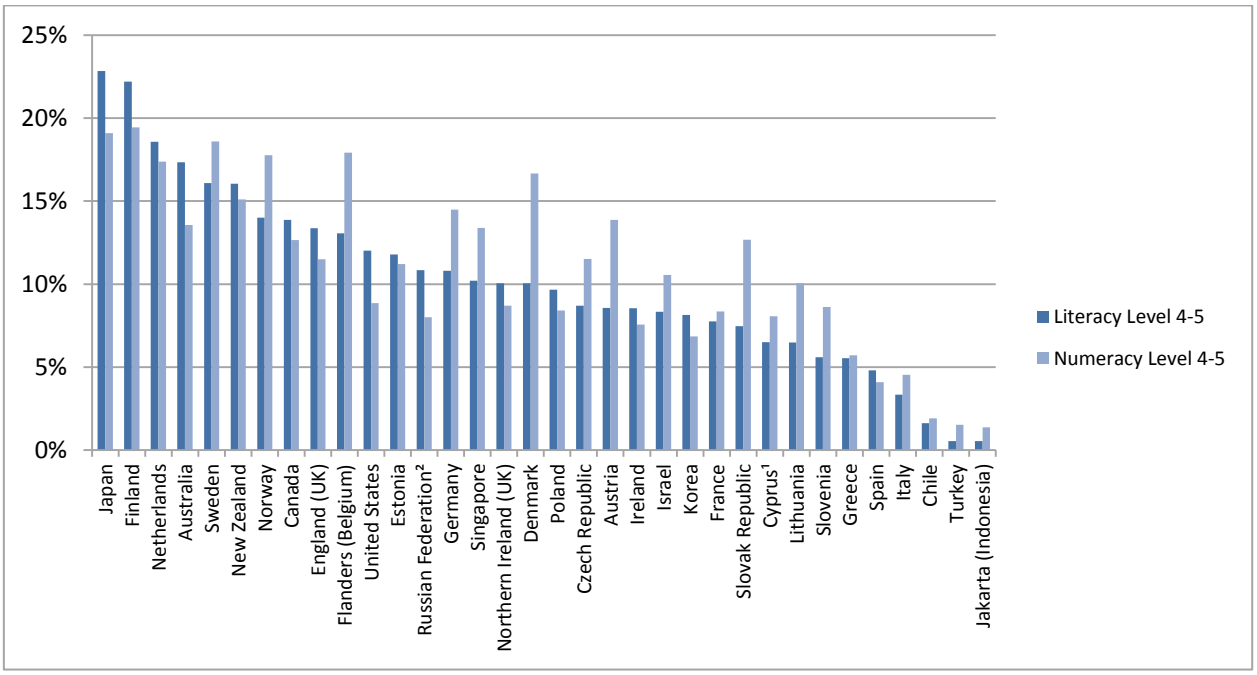
PIAAC assesses a set of general cognitive skills that are an important focus of development during education and that have been widely used at work. The findings on skill use demonstrate that large proportions of the workforce use these skills every day at work, even many workers with modest levels of proficiency. But these positive findings about skill use apply to today's economy at a time when many existing computer capabilities have not yet been broadly applied. The same conclusions will not necessarily hold as those capabilities begin to be generally available from computers. At that point, what levels of education and skill should we expect in the general population?

With respect to the skills included in PIAAC, there is not likely to be strong demand for human workers except for those whose proficiency levels are quite high. The expert projections suggest that in a decade or two workers will need proficiency in literacy and numeracy at Level 4 or 5 to be clearly better than computers in these general cognitive skills.[39] However, on average only 12% and 13% of the OECD workforce, respectively, has proficiency in literacy and numeracy at these levels. As a result, most of the workforce could not distinguish itself from computers in these skill areas.

[39] For the third skill area of problem solving with computers, the projected capabilities of computers are already close to the top of the scale.

One likely response to increasing computer capabilities would be to attempt to increase the level of skills in the workforce so that more people have skills that are greater than computer capabilities. Most countries around the world have worked to increase the education and skills of their populations and this strategy could have a number of beneficial effects. We do not yet see an increase in the proportion of workers at higher proficiency levels as a result of past education improvements - indeed, the analysis in Chapter 2 suggests there has been a modest decrease over the past two decades - but potentially this could change. By looking across countries, we can identify those that are more successful in achieving high proportions of adults with proficiencies in literacy and numeracy at Levels 4 and 5 and these examples indicate what improvements may be possible in other countries. Figure 5.6 shows the proportion of adults at the higher proficiency levels for all 34 countries and economies that have participated in PIAAC, including both OECD and non-OECD countries. The figure shows a wide range of results across the countries and suggests that many countries could substantially improve. However, the maximum - 23% for literacy and 19% for numeracy, both for Japan - is distinctly limited. The average performance of the best country suggests that only a quarter of the population could be better than projected computer capabilities in literacy and numeracy.

Figure 5.6 Proportion of adults with high literacy and numeracy proficiency, by country



Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A5.6.

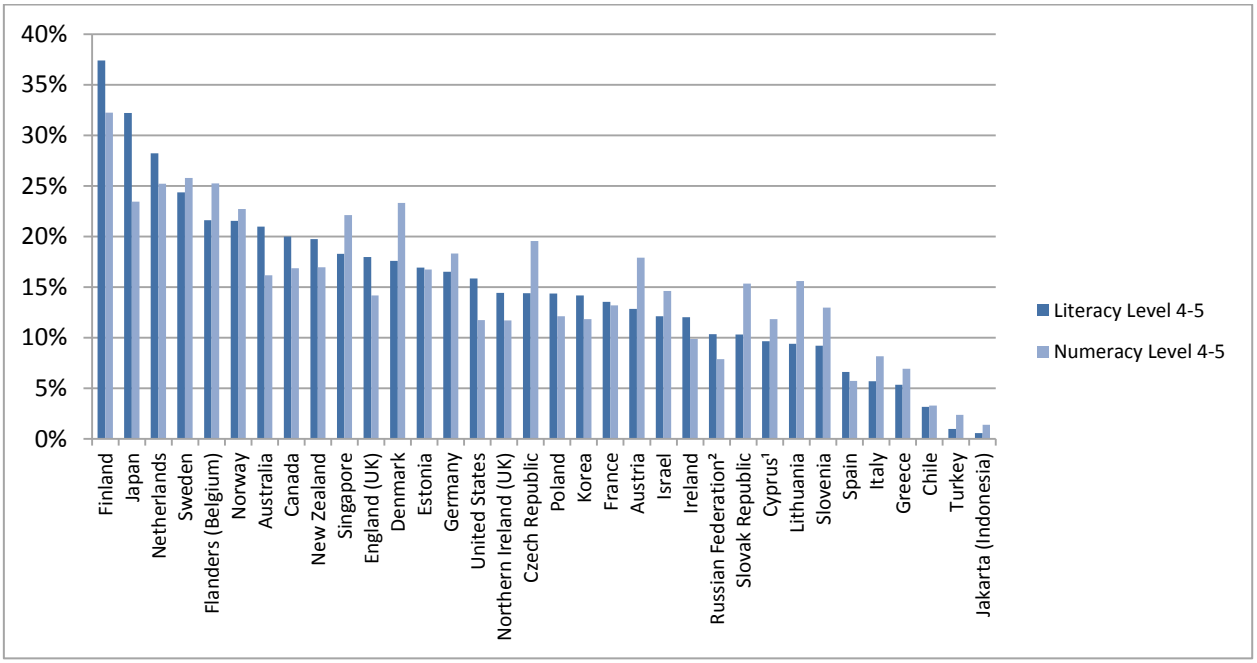
[1] See footnote 5.

[2] Readers should note that the sample for the Russian Federation does not include the population of the Moscow municipal area. The data published, therefore, do not represent the entire resident population aged 16-65 in Russia but rather the population of Russia excluding the population residing in the Moscow municipal area.

The proficiency of the full population presents a more pessimistic picture of full skill potential, since many older people received less education and less effective education than people who are educated today. In addition, the skills of older people may have weakened over time if they have not used them regularly. In general, PIAAC finds that skill levels are highest for the cohort of adults that has most recently fully completed formal education and declines for older cohorts. Figure 5.7 shows the results for the highest-performing cohort of adults, those aged 25-34. The OECD average is significantly higher for this group than for the full population, by 5 percentage points for literacy and 4 percentage points for

numeracy. However, in the highest-achieving country still only about a third of these younger adults reach the higher proficiency levels in literacy and numeracy - 37% for literacy and 32% for numeracy, both for Finland.

Figure 5.7 Proportion of adults aged 25-34 with high literacy or numeracy proficiency, by country



Source: Survey of Adult Skills (PIAAC) (2012, 2015), Table A5.7.

[1] See footnote 5.

[2] See note on previous page.

Variation across countries and across age cohorts suggests that many more workers could achieve proficiencies in literacy and numeracy at Levels 4 and 5. However, there is no indication in the performance of the highest-performing cohort in the highest-performing countries that a majority of the population could reach the higher levels of proficiency. Furthermore, even if increasing average proficiencies to the levels of the highest-performing cohorts and countries is possible, it would certainly take decades for other countries to achieve these results. And during that time computer capabilities in these skill areas will continue to improve.

With respect to general cognitive skills, higher levels of proficiency in literacy and numeracy are likely to be important for some part of the workforce over the next several decades as computer capabilities for the lower levels of literacy and numeracy are applied. However, it does not appear that these skills can be the key to employability for the majority of the workforce over this period. Given the levels of proficiency demonstrated in the past, it is simply not plausible that most workers over the next couple decades will be able to achieve higher levels of literacy and numeracy than available computer capabilities. If we want to understand what skills workers are likely to need over the next several decades, we need to know much more about the other kinds of skills that workers use beyond the general cognitive skills assessed by PIAAC. And we need to understand the levels of proficiency computers are developing with respect to these skills as well.

Assessing a broader range of skills for adults and computers

Research on job analysis in industrial and organizational psychology has resulted in several different approaches for understanding and categorizing work-related skills and tasks (Fleishman, Quaintance and Broedling, 1984; National Research Council, 2010). These taxonomies provide a way of systematically considering the range of skills used at work and the way these different skills are brought together in different kinds of tasks. Some of these skills - like literacy and numeracy - are developed during formal education, whereas others - like physical dexterity or social perception - are primarily developed outside of formal education. It is necessary to understand how all these skills come together at work to be able to understand how worker activities will change as new computer capabilities develop and how the education system should evolve in response.

Existing education assessments understandably focus on the skills that are developed during formal education. The OECD's PISA assessment of 15-year olds is an example of this type of test. Although PIAAC provides information about adults rather than students, it still focuses on skills primarily developed during education. Recent efforts to understand the importance of social and emotional skills make the case that they are affected by education and should be included in the set of education outcomes that are assessed as a part of education research and policy (OECD, 2015b). And PISA continually explores relevant new content domains, such as problem solving and financial literacy. However, there are many key work skills that are not considered in these testing programmes because they are not developed primarily during formal education.

Outside of formal education, there is a rich tradition of assessment of work-related skills used for occupational licensing and worker selection and training (e.g., Fleishman and Reilly, 1995; National Research Council, 1991, 2001, 2015). This work provides a set of tools that could be used to describe more precisely what skills workers need in different situations and how they relate to computer capabilities. The approach taken in this exploratory project provides a way to use such assessments to connect information about the skill proficiency of workers to the judgments of computer scientists about the growing capabilities of computers. To understand how computers will likely change the full range of skills used in the economy, this work should be extended across the full range of work skills.

At a time when computers are developing capabilities across a wide range of skill areas, policymakers need to have a much more systematic picture of work skills than is provided by tests of education-related skills alone. Because different skills are used together to perform work tasks, information about education-related skills alone cannot provide information even about those education-related skills themselves. The interdependence between different skills was readily demonstrated during the computer scientist review of the PIAAC literacy and numeracy questions, where one of the greatest issues in comparing human and computer capabilities in those two skill areas stemmed from the need for skills related to vision or common sense to answer many of the questions assessing literacy or numeracy. This skill interdependence in the context of the PIAAC test questions is merely an example of the interdependence that occurs throughout the workplace. We need information about the full set of skills to understand which skills workers will need in the future and how they are likely to interact with the capabilities that computers will increasingly be able to provide.

REFERENCES

Comin, D., and B. Hobijn, 2010, An Exploration of Technology Diffusion, *American Economic Review, 100*(5), 2031-59.

Fleishman, E.A., M.K. Quaintance, and L.A. Broedling, 1984, *Taxonomies of Human Performance: The Description of Human Tasks*, Orlando, Florida: Academic Press.

Fleishman, E.A., and M.E. Reilly, 1995, *Handbook of Human Abilities: Definitions, Measurements, and Job Task Requirements*, Manpower Research Institute.

Griliches, Z., 1957, Hybrid Corn: An Exploration in the Economics of Technological Change, *Econometrica, 25(4),* 501-522.

National Research Council, 2015, *Measuring Human Capabilities: An Agenda for Basic Research on the Assessment of Individual and Group Performance Potential for Military Accession*. Committee on Measuring Human Capabilities: Performance Potential of Individuals and Collectives. Washington, DC: The National Academies Press.

National Research Council, 2010, *A Database for a Changing Economy: Review of the Occupational Information Network (O*NET)*. Panel to Review the Occupational Information Network (O*NET), N.T. Tippins and M.L. Hilton, eds. Washington, DC: The National Academies Press.

National Research Council, 2001, *Testing Teacher Candidates: The Role of Licensure Tests in Improving Teacher Quality*. Committee on Assessment and Teacher Quality, K.J. Mitchell, D.Z. Robinson, B.S. Plake, and K.T. Knowles, eds. Washington, DC: The National Academies Press.

National Research Council, 1991, Performance Assessment for the Workplace: Volume 1. Committee on the Performance of Military Personnel, A.K. Wigdor and B.F. Green, Jr., eds. Washington, DC: The National Academies Press.

OECD, 2015a, *The Future of Productivity*, OECD Publishing.

OECD, 2015b, *Skills for Social Progress: The Power of Social and Emotional Skills*, OECD Skills Studies, OECD Publishing. Paris. http://dx.doi.org/10.1787/9789264226159-en.