OECD Social, Employment and Migration Working Papers
No. 314

Measuring disability employment gaps: How to get robust comparisons across countries and over time

Ben Geiger,

Christopher Prinz

https://dx.doi.org/10.1787/2d0be829-en

OECD

OECD Social, Employment and Migration Working Papers No.314

Measuring disability employment gaps

How to get robust comparisons across countries and over time

JEL classification: H55, I14, I18, J14, J18, K31, K38
Keywords: disability measurement, disability assessment, disability prevalence, disability employment, employment gaps

Authorised for publication by Stefano Scarpetta, Director, Directorate for Employment, Labour and Social Affairs.

Additional methodological details, including the full code for the analysis, additional technical explanations, and additional figures and tables, including confidence intervals, can be found in an online appendix under the link here.

Christopher Prinz, Ben Geiger, King's College London (Professor in Social Science and Health)
christopher.prinz@oecd.org; ben.geiger@kcl.ac.uk

Disclaimers

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Member countries of the OECD.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Note by the Republic of Türkiye
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. Türkiye recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Türkiye 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 Türkiye. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

Abstract

This Working Paper addresses a key issue for research on disability issues and policies: the volatility of the measurement of disability and its impact on indicators based on this measure, such as employment rates and employment gaps. People in different countries and in the same country at different times differ in how likely they are to report a disability. This makes it difficult if not impossible to measure the impact of policies on disability employment outcomes. Even worse, a successful policy may change the way that people report disability, in ways that make the disability employment gap look worse – even though the policy has been a success. To overcome these issues, this report discusses different ways of producing more reliable estimates of disability employment rates and gaps and trends over time, to complement the disability employment gap indicator.

Acknowledgements

SHARE data collection has been primarily funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812) and FP7 (SHARE-PREP: N°211909, SHARE-LEAP: N°227822, SHARE M4: N°261982). Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

ELSA is sponsored by the US National Institute of Aging, and UK Department of Health, Department for Work and Pensions, Office for National Statistics, Department for Transport, HM Revenue and Customs, and Department for Communities and Local Government.

HRS is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.

Gateway to Global Aging Data is produced by the USC Program on Global Aging, Health & Policy, with funding from the National Institute on Aging (R01 AG030153).

This report is based on data from Eurostat (the European Health Interview Survey, EHIS, waves 2 and 3). The responsibility for all conclusions drawn from the data lies entirely with the authors.

Helpful comments to an earlier version of the paper were provided by Jessica Mahoney and Carry Exton from the OECD Centre on Wellbeing, Inclusion, Sustainability and Equal Opportunity, Mark Keese, Mark Pearson and Andrea Bassanini from the OECD's Employment, Labour and Social Affairs Directorate, Philip Hemmings from the OECD's Economics Directorate, and delegates from several OECD member countries during and following the discussion of the paper with the Working Party on Employment. Special thanks go to Aleksandra Posarac (former Lead Economist, World Bank) and Prof. Jerôme Bickenbach (Disability Policy Lead, Swiss Paraplegic Research) for extensive discussions and suggestions. Michael van den Berg from the OECD's Health Division prepared the material for the box on the PaRIS survey, Dana Blumin helped with some of the charts, and Monica Meza-Essid helped with the final formatting of the paper.

Table of contents

Executive summary

The disability employment gap – the percentage-point difference in employment rates between people with and without disability – is a core indicator of disability inclusion. It is obtainable from population survey data and provides an overall assessment of the extent to which people with disability are included (or excluded) from the labour market in each country. As such, it has been regularly used by the OECD, the European Commission, national governments, and increasingly also by academics.

However, this indicator, in the rest of this paper referred to as the "conventional disability employment gap", may not be fully accurate. Most measures of disability inclusion rely on simple measures of disability prevalence, where people report whether they have a health condition that affects their day-to-day activities ('single-item activity-limiting disability'). These measures have some value but are poorly comparable both across countries and within countries over time, because they are affected by people's environment. Disability is intrinsically about the interaction of a person's impairments with their environments, but we cannot have a disability measure that itself already factors in the role of policies, because this makes it impossible to tease apart how policies affect the inclusion of people with disability. Indeed, because it changes in response to policies, there are times that the conventional disability employment gap can make success look like failure, which does not provide a robust basis for policymaking.

In response, this report reviews different ways of producing estimates of the disability employment gap. There is wide agreement that more detailed surveys covering multiple different aspects of disability are needed, but disagreements about the best way of doing this. The report therefore critically reviews the latest methodologies for measuring disability in the wider literature, proposes some new methods, which are explained conceptually, and outlines how to implement them in practice (and makes a sample code available for replication). Finally, these proposed new methods are used to compare which countries perform best in minimising disability employment gaps.

Given that single-item disability measures are common, Chapter 2 of the paper reviews how to use these measures to look at disability employment rates and gaps as robustly as possible. It recommends that researchers and policymakers…

- …use *more-comparable surveys* using input harmonization. Coordinated national studies are invaluable for showing that people with disability face employment barriers, and providing large, high-quality datasets for research. But for robust comparisons of countries, the flexibility that is necessary for coordinated national studies means we cannot be sure if the differences we see reflect genuine patterns or the results of varying methodological choices. International studies that harmonize the methods of data collection are therefore the best source for robust comparisons.
- …use a further measure alongside the conventional disability employment gap, the *prevalence-adjusted disability employment gap*. This is the conventional disability employment gap multiplied by the prevalence of disability (and, therefore, shows the share of the population potentially prevented from working due to disability). Despite being simple, it can still make comparisons over time and place more robust. Sometimes this is because the prevalence-adjusted gap is more likely to be unbiased. More commonly, however, similar patterns in the conventional and the prevalence-adjusted gap measures can give more confidence in the results. Chapter 2 ends by demonstrating the use of the proposed prevalence-adjusted gap in two case studies; one comparing European countries, the other looking at one country, the United Kingdom, over time.

The takeaway message is that the conventional disability employment gap and the prevalence-adjusted gap are useful complements to each other – they highlight different things that policymakers need to know; and where they show different patterns, they alert us to the possibility that we cannot necessarily trust these single-item measures of activity-limiting disability.

However, the most robust comparisons require better measures of disability. It is widely recognised that the best way of doing this is to use **a series of specific questions on impairments and activity limitations**, which are answered reasonably similarly wherever and whenever they are asked, which are then **combined into a single scale**. The resulting disability measure does not refer to people with impairments who are *actually disadvantaged* given their particular environment but instead refers to people with impairments who are *potentially disadvantaged* in the environments we see across OECD countries. It is important to follow the WHO's International Classification of Functioning, Disability and Health (ICF) framework when interpreting these measures, but to uncover disabling social environments, we need an underlying measure of disability prevalence that is measured consistently in different times and places.

While there are many examples of creating disability scales, researchers often seem unaware of the limitations of their methods, and the different choices available that might produce more robust results. In Chapter 3, the paper recommends that:

- *The items included in the scale should be comparable and comprehensive:* they should focus on measures that (i) are likely to be comparable across time and place; and (ii) cover all dimensions of impairments/activity limitations that are likely to be work-disabling in OECD countries.
- *The weights for each item in the scale should be produced using the 'predicted disability' approach* – that is, by weighting each item of the scale according to how strongly it predicts single-item activity-limiting disability.
- *The scale should be turned into a binary disability category using a probabilistic approach*, rather than a fixed threshold. This has not been used previously but is likely to be more comparable across time and place, and to be less sensitive to arbitrarily fixed thresholds for disability.

Previous studies have not taken seriously comparative differences in disability employment gaps, because it has been assumed that these are methodologically unsound. When using more robust (albeit imperfect) methods, it seems that there are two cross-national patterns that can helpfully drive future research, which may in turn help improve inclusion for people with disability into work: (1) Why do countries with lower overall employment rates often have lower disability employment gaps? and (2) How do some countries buck this trend, combining high disability employment rates with low disability employment gaps?

Unfortunately, the possibility of using the proposed more-robust disability measures is constrained by the limited availability of detailed comparative data on impairments and activity limitations. This is not because validated question sets do not exist – several have been developed, including the WHO Model Disability Survey – but because these have not been used in an input-harmonized survey across OECD countries. Because of the length of these question sets, the most cost-effective approach is to supplement existing surveys with periodic 'calibration surveys' that provide more robust estimates of both disability and employment. The paper strongly encourages researchers, policy makers and statistical agencies to work together to create such calibration surveys, to provide better data for international comparisons of disability outcomes across the full working-age population.

Finally, whatever techniques are used, researchers need to be aware of the challenges of comparability in measures of the disability employment gap. International comparisons and comparisons within countries over time are both difficult and crucial for policy making; they need to be done with care.

Résumé

L'écart d'emploi lié au handicap – la différence en points de pourcentage entre les taux d'emploi des personnes handicapées et celles qui ne le sont pas – est un indicateur clé de l'inclusion du handicap. Il est obtenu à partir de données d'enquêtes de population et fournit une évaluation globale du degré d'inclusion (ou d'exclusion) des personnes handicapées du marché du travail dans chaque pays. À ce titre, il est régulièrement utilisé par l'OCDE, la Commission européenne, les gouvernements nationaux et, de plus en plus, par le monde universitaire.

Toutefois, cet indicateur, appelé dans la suite de cet article « écart conventionnel d'emploi lié au handicap », peut ne pas être entièrement exact. La plupart des mesures de l'inclusion du handicap reposent sur des mesures simples de la prévalence du handicap, où les personnes déclarent si elles souffrent d'un problème de santé affectant leurs activités quotidiennes (« incapacité limitant l'activité d'un seul élément »). Ces mesures ont une certaine valeur, mais sont difficilement comparables entre les pays et au sein d'un même pays au fil du temps, car elles sont influencées par l'environnement des personnes. Le handicap est intrinsèquement lié à l'interaction des déficiences d'une personne avec son environnement. Cependant, il est impossible d'avoir une mesure du handicap qui prenne déjà en compte le rôle des politiques, car cela rend impossible de distinguer l'impact de ces dernières sur l'inclusion des personnes handicapées. En effet, évoluant en fonction des politiques, l'écart d'emploi conventionnel lié au handicap peut parfois faire passer la réussite pour un échec, ce qui ne constitue pas une base solide pour l'élaboration des politiques.

En réponse, ce rapport examine différentes manières d'améliorer les estimations de l'écart d'emploi lié au handicap. Il existe un large consensus sur la nécessité d'enquêtes plus détaillées couvrant de multiples aspects du handicap, mais des désaccords subsistent quant à la meilleure méthode. Le rapport examine donc de manière critique les dernières méthodologies de mesure du handicap dans la littérature générale, propose de nouvelles méthodes, expliquées conceptuellement, et décrit leur mise en œuvre pratique (et met à disposition un exemple de code pour la réplication). Enfin, ces nouvelles méthodes proposées sont utilisées pour comparer les pays les plus performants en matière de réduction des écarts d'emploi liés au handicap.

Étant donné la fréquence des mesures du handicap par item unique, le chapitre 2 de l'article examine comment utiliser ces mesures pour analyser les taux et les écarts d'emploi liés au handicap de la manière la plus fiable possible. Il recommande aux chercheurs et aux décideurs politiques…

- …d'utiliser des enquêtes plus comparables utilisant l'harmonisation des données d'entrée. Les études nationales coordonnées sont précieuses pour démontrer que les personnes handicapées rencontrent des obstacles à l'emploi et fournir des ensembles de données vastes et de qualité pour la recherche. Cependant, pour des comparaisons fiables entre pays, la flexibilité nécessaire aux études nationales coordonnées ne permet pas de savoir avec certitude si les différences observées reflètent de véritables tendances ou le résultat de choix méthodologiques différents. Les études internationales harmonisant les méthodes de collecte de données constituent donc la meilleure source pour des comparaisons fiables.
- …d'utiliser une mesure supplémentaire, en plus de l'écart d'emploi lié au handicap conventionnel, l'écart d'emploi lié au handicap ajusté en fonction de la prévalence. Il s'agit de l'écart d'emploi lié au handicap conventionnel multiplié par la prévalence du handicap (et, par conséquent, indique la part de la population potentiellement empêchée de travailler en raison du handicap). Malgré sa

simplicité, cet indicateur permet de renforcer la robustesse des comparaisons dans le temps et dans l'espace. Parfois, cela s'explique par le fait que l'écart ajusté en fonction de la prévalence est plus susceptible d'être impartial. Plus généralement, cependant, des tendances similaires dans les mesures conventionnelles et l'écart ajusté en fonction de la prévalence peuvent renforcer la fiabilité des résultats. Le chapitre 2 se termine par une démonstration de l'utilisation de l'écart ajusté en fonction de la prévalence proposé dans deux études de cas : l'une comparant des pays européens, l'autre un pays, le Royaume-Uni, au fil du temps.

Le message à retenir est que l'écart conventionnel d'emploi lié au handicap et l'écart ajusté en fonction de la prévalence se complètent utilement : ils mettent en évidence des éléments différents que les décideurs politiques doivent connaître ; et lorsqu'ils présentent des tendances différentes, ils nous alertent sur le fait que nous ne pouvons pas nécessairement nous fier à ces mesures mono-items de l'incapacité limitant les activités.

Cependant, les comparaisons les plus solides nécessitent de meilleures mesures de l'incapacité. Il est largement reconnu que la meilleure façon d'y parvenir est d'utiliser une série de questions spécifiques sur les déficiences et les limitations d'activité, auxquelles les réponses sont relativement similaires où et quand elles sont posées, puis combinées dans une seule échelle. La mesure du handicap qui en résulte ne fait pas référence aux personnes handicapées réellement défavorisées compte tenu de leur environnement particulier, mais plutôt aux personnes handicapées potentiellement défavorisées dans les environnements observés dans les pays de l'OCDE. Il est important de suivre le cadre de la Classification internationale du fonctionnement, du handicap et de la santé (CIF) de l'OMS pour interpréter ces mesures. Cependant, pour identifier les environnements sociaux invalidants, il est nécessaire de disposer d'une mesure sous-jacente de la prévalence du handicap, mesurée de manière cohérente à différentes époques et dans différents lieux.

Bien qu'il existe de nombreux exemples de création d'échelles de handicap, les chercheurs semblent souvent ignorer les limites de leurs méthodes et les différents choix possibles qui pourraient produire des résultats plus fiables. Au chapitre 3, l'article recommande que:

- Les éléments de l'échelle soient comparables et exhaustifs : ils doivent se concentrer sur des mesures qui (i) sont susceptibles d'être comparables dans le temps et dans l'espace ; et (ii) couvrent toutes les dimensions des déficiences/limitations d'activité susceptibles d'entraîner une incapacité professionnelle dans les pays de l'OCDE.
- La pondération de chaque élément de limitation fonctionnelle de l'échelle doit être calculée selon l'approche de « prédiction du handicap », c'est-à-dire en pondérant chaque élément de l'échelle en fonction de sa capacité à prédire un handicap limitant l'activité pour un seul élément.
- L'échelle devrait être transformée en une catégorie binaire de handicap utilisant une approche probabiliste, plutôt qu'un seuil fixe. Cette approche n'a jamais été utilisée auparavant, mais elle est susceptible d'être plus comparable dans le temps et l'espace, et d'être moins sensible aux seuils d'invalidité fixés arbitrairement.

Les études antérieures n'ont pas pris au sérieux les différences comparatives dans les écarts d'emploi liés au handicap, car on a supposé qu'elles étaient méthodologiquement peu fiables. L'utilisation de méthodes plus robustes (bien qu'imparfaites) révèle deux tendances transnationales qui pourraient orienter les recherches futures, ce qui pourrait à son tour contribuer à améliorer l'inclusion des personnes handicapées dans l'emploi : (1) Pourquoi les pays ayant des taux d'emploi globaux plus faibles ont-ils souvent des écarts d'emploi liés au handicap plus faibles ? et (2) Comment certains pays parviennent-ils à inverser cette tendance, en combinant des taux d'emploi liés au handicap élevés et de faibles écarts d'emploi liés au handicap?

Malheureusement, la possibilité d'utiliser les mesures d'invalidité plus robustes proposées est limitée par la disponibilité limitée de données comparatives détaillées sur les limitations fonctionnelles. Cela ne

s'explique pas par l'absence de questionnaires validés – plusieurs ont été élaborés, notamment l'Enquête modèle de l'OMS sur le handicap – mais par le fait qu'ils n'ont pas été utilisés dans une enquête harmonisée des données d'entrée dans les pays de l'OCDE. Compte tenu de la longueur de ces questionnaires, l'approche la plus rentable consiste à compléter les enquêtes existantes par des « enquêtes d'étalonnage » périodiques fournissant des estimations plus fiables du handicap et de l'emploi. Ce document encourage vivement les chercheurs, les décideurs politiques et les agences statistiques à collaborer à la création de telles enquêtes d'étalonnage, afin de fournir de meilleures données pour les comparaisons internationales des résultats en matière de handicap sur l'ensemble de la population en âge de travailler.

Enfin, quelles que soient les techniques utilisées, les chercheurs doivent être conscients des défis liés à la comparabilité des mesures de l'écart d'emploi lié au handicap. Les comparaisons internationales et intra-pays au fil du temps sont à la fois difficiles et cruciales pour l'élaboration des politiques ; elles doivent être réalisées avec prudence.

1 Why better methods for measuring disability employment gaps are needed

The disability employment gap is a core indicator of disability inclusion. However, the conventional disability employment gap may not be fully accurate, as the definition of disability already factors in the role of policies, making it impossible to tease apart how such policies affect disability inclusion. What is worse, a successful policy may change the way that people report disability, in ways that make the disability employment gap look worse – even though the policy has really been a success.

In response, this paper reviews different ways of producing better estimates of the disability employment gap, to complement the conventional disability employment gap indicator. It critically reviews latest methodologies for measuring disability in the wider literature, proposes new methods, which are explained conceptually, and outlines how to implement them in practice. The paper also uses these methods to look substantively at which countries seem to perform best in minimising disability employment gaps.

The conventional disability employment gap – the difference in the employment rates of people with vs. without disability[1] – is a core indicator of disability inclusion in OECD countries. Arguably it is not possible or appropriate for all people with disability to work; but it is important to see which countries do a better job of including people with disability in the labour market. The conventional disability employment gap shows this, and it is easily obtainable from survey data. As such, it has formed a core part of OECD reports on disability policy (OECD, 2003, 2010a, 2012, 2022), as well as being used in European Commission policymaking (Grammenos, 2025; Priestley and Grammenos, 2021), including in the EU's social scoreboard. It is also increasingly used in academic papers (Geiger *et al.*, 2019; Geiger *et al.*, 2017; Gugushvili *et al.*, 2023; Reinders Folmer *et al.*, 2020; van der Zwan and de Beer, 2021).

However, the conventional disability employment gap may be highly misleading. Similar people in different countries, or at different times in the same country, differ in how likely they are to report a disability. This is likely to bias the disability employment gap, possibly in self-defeating ways, as better inclusion may make disability employment gaps look worse. This is because people with less severe limitations may become less likely to report a disability, leaving the indicator reflecting only those people with more severe limitations, who are less likely to be employed. If an indicator can make success look like failure, then it does not provide a robust basis for policymaking.

In response, this report discusses how to produce better estimates of the disability employment gap. There is wide agreement that more detailed surveys covering multiple different aspects of disability are needed, but disagreements about the best way of doing this. The report therefore critically reviews the latest methodologies for measuring disability in the wider literature, including influential methods proposed by the National Bureau of Economic Research (Poterba *et al.*, 2013), the World Disability Report (WHO, 2011), and the World Health Organization (WHO) more recently (Sabariego *et al.*, 2022). It also proposes new methods, explains them conceptually and outlines how to implement them in practice. Throughout the report, it produces new estimates of the disability employment gap using these different methods.

The report has four main chapters. Chapter 2 looks at the disability employment gap using single-item measures of activity-limiting disability. Chapter 3 looks at the gap using multi-item scales that combine information on multiple aspects of ill-health and impairment. Chapter 4 uses these new measures to see which countries seem to be doing best in minimising disability employment gaps, and what this means for policy. Finally, Chapter 5 summarises the discussion and resulting recommendations for policymakers, data producers and researchers. Before this, however, the rest of Chapter 1 sets out what the authors mean by 'disability', and why it is so difficult to compare across countries.

Understanding 'disability'

To measure disability employment gaps robustly, we need to start by having a clear idea of what 'disability' is. The most widely used international model is the WHO's International Classification of Functioning, Disability and Health ('ICF'), which distinguishes between:

- *impairments* (problems in body functions and body structures);
- *activity limitations* (difficulties executing a task); and
- *participation restrictions* (a problem experienced in a life situation, e.g., working).

Crucially, the ICF distinguishes between *'capacity'* (or 'biological health') and *'performance'* (what people can actually do in the environment that they are in) (Bickenbach *et al.*, 2023). [2] Disability cannot be understood without looking at context, because people with the same intrinsic health capacity will perform differently depending on their physical, built and social environments. Given our focus on the workplace, this means that a person with impairments may not have participation restrictions in a country with an inclusive labour market but face restricted participation in a country with worse working conditions.

We use these terms throughout the report and return to discuss the ICF in more detail in Chapter 3.

Why the conventional disability employment gap may not be fully accurate

The simplest and most common way of measuring disability (at least in OECD countries) is to ask people directly if they experience activity limitations/participation restrictions. This is usually done with one or two questions that ask people how much their health restricts their daily life (which we term 'activity-limiting disability questions'). Extensive efforts have been made to create such a measure that is internationally comparable, such as the EU's Global Activity Limitation Indicator (GALI), which asks about whether people are limited in the *'activities people usually do'* (Robine and Jagger, 2003).

For some purposes, these measures are a sensible choice – they are simple and focus on people whose participation in society is restricted because of the interaction between their impairments and their environment, which is the group that policymakers are usually concerned about. They are also better than asking people if they consider themselves to be disabled, which misses out many people with restricted participation who would nevertheless not use the term 'disability' to describe themselves. Yet, these activity-limiting disability questions are problematic for other purposes – including international comparisons, time trends, or policy evaluation.

Put simply, people with a given set of impairments vary a lot in whether they report an activity-limiting disability. This is partly because the question is quite vague, leaving a lot of space for individual/cultural variation in how people respond. But a bigger problem is that whether people say that something is activity-limiting depends on the environment they are in (or in the ICF terms above: these questions refer to real-world performance rather than intrinsic capacity). As such, they depend both on individually-variable environments, affected by socioeconomic status/wealth and social support (Berger *et al.*, 2015), and on (sub-)national disability-related policies themselves. This is desirable in some respects – disability is inherently about the interaction of impairments with environments. But this makes these questions unsuitable for monitoring policy success or failure – policies will not just affect the employment rates of people with disabilities, but they will affect whether people report an activity-limiting disability *per se*.

On paper, these biases probably make countries appear to be less successful if they have a stronger policy focus on disability. For example, compared to other OECD countries, we may expect people with a given capacity in Norway to be both:

- *…more likely to report an activity-limiting disability*, because the Norwegian state classifies so many people as having a disability within the social protection system.[3] Because the state has told them they have reduced work capacity, Norwegians may be more likely to say that their limitations affect their day-to-day activities (as found in vignettes, e.g., Yin and Heiland, 2015).
- *…less likely to report an activity-limiting disability*, because the state makes greater efforts to include people with disability in the workplace and society compared to many other countries (Mont, 2007:4.4, 4.11). Because people experience fewer barriers in everyday life, Norwegians may be less likely to say that their limitations affect their day-to-day activities.

These two underlying mechanisms are *both* likely to work in the same direction, biasing the conventional disability employment gap so that it appears worse in countries like Norway than elsewhere.[4] If Norway has a high disability employment gap compared to other countries, then it is unclear whether this is because its policies are genuinely ineffective, or simply because these policies change the way that people report activity-limiting disability.

Conversely, existing studies have suggested that some countries with more limited disability inclusion policies nevertheless seem to perform well (e.g., Italy). Again, this may be because of biases around the reporting of disability in those countries compared to other countries (as we have just described) – or it may be because there are previously unrecognized forces at play that are effective in including people with

disability in the labour market, or because of a low employment rate overall in general, or other factors (OECD, 2023). Without the use of a robust disability measure that we can trust, we cannot really use the disability employment gap as a source of insight into what works in disability inclusion.

Conventional disability measures have their place

It is important to note that while conventional disability measures such as GALI have problems, they nevertheless have continued value, for several reasons:

- *They are effective in showing that people with disability face employment barriers in all countries.* They are therefore valuable for drawing attention to disability-related exclusion within and across OECD countries (Priestley and Grammenos, 2021) and the need for action (Fehr *et al.*, 2017);
- *They make visible the employment situation of a politically important group* – that is, people who *perceive* that their health or impairments limit the activities they can do. It is helpful to show the situation of people so-defined, as long as this is not interpreted as the true labour market situation of 'consistently measured' people with disability;
- *They are much cheaper and easier to collect data on.* As the report explains below, alternatives to single-item activity-limiting disability often require a greater (sometimes much greater) number of questions, which is often implausible in multi-purpose surveys such as Labour Force Surveys and will increase data collection costs even where it can be included.

In the rest of the report, we pay attention to both the continued value of single-item activity-limiting disability and the trade-offs between different measures – concluding with recommendations about how to efficiently get robust knowledge using a *combination* of conventional and newer measures.

An outline of this report

The conventional disability employment gap (and the general activity-limiting disability measure on which it is based) do have value, but they are problematic for evaluating the success or failure of policy. In the rest of this report, we review a series of different choices available to policymakers, data producers and researchers who want to produce more robust measures of the disability employment gap and its development over time.

We firstly look at ways of using single self-reported measures of disability, focusing on two choices:

- *Choosing comparable data* collected in identical ways in different times and different places.
- *Accounting for the varying prevalence of disability*, by using a simple new measure that multiples the disability employment gap by the prevalence of disability.

We then look at ways of estimating the disability employment gap that use multi-item scales, putting together information on multiple aspects of impairments and activity limitations. We focus on three choices:

- *Choosing indicators for the scale,* and the importance of excluding self-reported health and medically diagnosed conditions that are less likely to be comparable across time and place;
- *How to create a disability scale,* comparing weights based on latent variable techniques (which are commonly used) with weights based on how well each item predicts disability;
- *Whether to turn the index into a binary measure of disability, and if so, how*.

The last two chapters look at which countries do a good job in minimising the disability employment gap and draw together the lessons from the methodological toolkit offered in this paper.

2 Measuring the disability employment gap using single-item measures

Given that single-item activity-limiting disability measures are common, we firstly review how to use these measures to look at disability prevalence and employment as robustly as possible. We recommend that researchers and policymakers (1) use more-comparable surveys, and (2) use a further measure alongside the conventional disability employment gap, the 'prevalence-adjusted disability employment gap'. This is the conventional disability employment gap multiplied by the prevalence of disability (and therefore shows the share of the population potentially prevented from working due to disability). Despite its simplicity, it can make comparisons over time and across countries more robust. The chapter ends by demonstrating the use of the prevalence-adjusted gap in two case studies, one comparing the disability employment gap across selected European countries, the other looking at the United Kingdom as a case study of the recent apparent weaknesses of the conventional disability employment gap.

In this chapter, we look at how to measure the disability employment gap using survey-based single-item activity-limiting disability measures, knowing that there is a risk that people will report these disability measures differently across different times and places. Given that this information is much cheaper to collect than the alternatives in Chapter 3 and the existing data are much more widely available, we here consider two ways of improving the robustness of conventional disability employment gap measures.

Choosing comparable data

Before considering measures and analytical approaches, it is important to start by considering the comparability of the data being used. We know that methodological choices may affect both disability prevalence and resulting social and labour market outcomes for people with disability – whether that is choices about survey mode (Bowling, 2005; Cernat *et al.*, 2016; Croezen *et al.*, 2016; Hood *et al.*, 2012), if someone reports for themselves or by proxy (Elkasabi, 2020; Lee *et al.*, 2004; Todorov and Kirchner, 2000), or the sampling approach and non-response biases (Croezen *et al.*, 2016; Groves and Peytcheva, 2008; Korkeila *et al.*, 2001). If we are going to compare disability across time and place, then we need to make sure that the methods used are as consistent as possible.

In practice, our choice is often between two different sources of comparative survey data on health and disability that contain different trade-offs. Firstly, there are surveys independently run by National Statistical Offices as part of coordinated research efforts ('coordinated national studies'). This includes surveys mandated by EU Regulations, such as the European Union Statistics on Income and Living Conditions (EU-SILC),[5] the European Health Interview Survey (EHIS) and the European Labour Force Survey (EULFS). Secondly, there are comparative surveys run by an international team ('international studies'), such as the European Social Survey (ESS) or the Global Aging Surveys (including the Survey of Health, Ageing and Retirement in Europe; see below).

Coordinated national studies have several strengths. They tend to be based on major Government surveys, which means they often have high sample sizes, frequent data collection, require little additional funding, have a prominent public role, and (in some countries) allow survey data to be combined with administrative data (Cases, 2021), as well as being part of long-running official statistical series. However, because these surveys have multiple purposes rather than just international comparisons, these coordinated efforts must allow some flexibility in their methodological choices (Cases, 2021). Often these surveys are therefore 'ex ante output-harmonized' (Wirth and Pforr, 2022) – that is, countries agree to produce data that contains information on certain topics, but can choose their preferred way of doing this (Arora *et al.*, 2015). As a result in, e.g., EU-SILC, countries differ in question wording, survey mode, use of proxy interviews, and sampling methods (Eurostat, 2011) – sometimes dramatically; the use of proxy interviews in 2009 varied from 1% in Sweden to 48% in Denmark (Eurostat, 2010).

International studies, in contrast, are usually 'input-harmonized' (Wirth and Pforr, 2022) – that is, the survey team tries to harmonise the data collection methods themselves, rather than just the outputs. For example, the European Social Survey (as used in Geiger *et al.*, 2017; Gugushvili *et al.*, 2023) won the prestigious Descartes Prize in 2005 for its efforts to reduce methodological variations in comparative surveys. Comparability is never perfect in international surveys – partly because of differences in, e.g., response rates and sampling frames, partly because there will always be cultural and institutional differences that influence question interpretation (Jowell, 1998). But surveys like the ESS reduce the extent of this, so that the results are much more likely to reflect genuine differences in disability prevalence and employment rather than methodological differences. The lack of flexibility comes at a cost, though, as such surveys tend to be smaller, less frequent, not as easily linked to administrative data, and not part of long-running official statistical series.

It must be emphasized that considerable efforts have been made to improve the comparability of coordinated national studies. For example, the Global Activity Limitations Indicator in Europe ('GALI')

('GALI'; see Robine and Jagger, 2003) has been adopted as a standard general disability measure (it asks about longstanding 'limitations because of health problems in activities people usually do' – a variant of the common 'limiting longstanding illness' questions used in many countries). There have also been efforts to increase the harmonization of methods in European surveys, e.g. in the Integrated European Social Statistics framework regulation adopted in 2019 (Cases, 2021). However, while this has moved towards input harmonization for the LFS, other surveys like EU-SILC still allow individual countries leeway in, e.g., survey mode or the use of administrative data. As Cases (2021) puts it *"while [the regulation] will undoubtedly contribute to improving comparability, it will not be a radical and definitive solution. Rather, it represents a significant step towards a process of convergence – a process that remains incomplete, despite achieving greater harmonisation through "inputs" (variables, content) than previously."*

In their current form, coordinated national studies are therefore less comparable than international studies – despite their wider value. They are still essential for showing that people with disability face employment barriers in all countries, using prominent national surveys and recent data (see Chapter 1). In the case of studies like EU-SILC, they provide large, high-quality, household-based and longitudinal data that has been extensively used for social science research, particularly on income and poverty (OECD, 2022; Wirth and Pforr, 2022). If the methods used are consistent within-countries over time (see discussion in Chapter 4), they provide crucial and regularly-updated pictures of trends in disability employment gaps within countries – although even here, we see large variation over time in some EU countries in disability prevalence measured by e.g. EU-SILC, suggesting that there are methodological issues affecting within-country analyses of trends. For robust comparisons of countries, though, the flexibility that is necessary for coordinated national studies means that we cannot be sure if the differences that we see reflect genuine patterns or the results of varying methodological choices.

Illustrating the impact of more comparable data

To illustrate this, we here estimate disability employment gaps using two surveys:

- *An output-harmonized survey,* the European Health Interview Survey, which is coordinated via an EU Regulation and also provides a model questionnaire that is followed by most countries. EHIS is one of the largest international health-focused surveys and central to Eurostat's attempts to collect comparable disability data in Europe.
- *An input-harmonized survey*, the Survey of Health, Ageing and Retirement in Europe (SHARE) (SHARE; Börsch-Supan, 2017), which is coordinated by a single team. It is also part of a wider network of Global Aging Surveys that use mostly identical question wording and methods, based on the US Health & Retirement Study (see https://g2aging.org/).

These surveys are chosen for this report as they can also be used in Chapter 3, as they both also include more detailed questions on health and disability. (The Global Aging Surveys have been the most common choice for researchers trying to use multi-item scales to compare disability across countries (e.g. Böheim and Leoni, 2015; Cieza *et al.*, 2015; Jürges, 2007; Wise, 2017)). Reflecting the trade-offs inherent in the choice of data source, it is important to note that SHARE is a survey of older people (aged 50+), so in much of Chapters 2 and 3 we focus on older working-age people (aged 50-69) in both surveys (for all-age EHIS results, see Chapter 4).

We use SHARE data from 2013-15, and EHIS data for 2014-15, using the supplied cross-sectional weights, and adjusting for age and gender to account for different demographic profiles in different countries (see Appendix A2). (We also take further steps to improve comparability; further methodological details are given in Appendix D, and the full code of the analyses is provided online alongside this report). We should note that the employment measures are slightly different in the two surveys, with EHIS (like EU-SILC) asking about self-defined employment status, while SHARE asks about working within the last month – an issue in other comparisons (Priestley and Grammenos, 2021:217). The disability measures, though, are meant to be the same.

The prevalence of disability in the two surveys is shown in Table 2.1. In many countries the prevalence of disability is similar, but in some countries, there are very large differences. These cannot be explained by random error; the significance of the differences between the surveys is p<0.0001. In Germany, for example, 45.5% of people aged 50-69 report a disability in SHARE, but only 28.1% report a disability in EHIS. Some of these are likely to be explained by differences in survey mode: SHARE is consistently face-to-face, while EHIS is online/postal in some countries (including Germany). Yet this is not a full explanation: for example, Denmark has the same differences in survey mode as Germany but shows a much *lower* (rather than higher) prevalence of disability in EHIS, compared with SHARE.

Table 2.1. Survey differences in the prevalence of single-item activity-limiting disability in 50-69-year-olds

	SHARE		EHIS		Difference	
Greece	18.2	[16.7, 19.7]	36.4	[34.1, 38.6]	-18.2	[-20.9, -15.5]
Italy	29.2	[27.2, 31.3]	28.6	[27.5, 29.7]	0.6	[-1.7, 3.0]
Denmark	30.7	[28.9, 32.6]	40	[37.7, 42.3]	-9.3	[-12.2, -6.3]
Sweden	32	[29.9, 34.1]	31.2	[28.2, 34.3]	0.7	[-2.9, 4.4]
France	33.7	[31.6, 35.8]	31.4	[29.8, 33.0]	2.3	[-0.3, 5.0]
Austria	39.8	[37.4, 42.2]	41.5	[39.8, 43.2]	-1.7	[-4.6, 1.2]
Czechia	40.8	[37.4, 44.1]	43.6	[41.2, 45.9]	-2.8	[-6.9, 1.3]
Slovenia	41.2	[38.8, 43.6]	46.8	[44.3, 49.3]	-5.6	[-9.1, -2.2]
Germany	45.5	[43.7, 47.3]	28.1	[26.8, 29.4]	17.4	[15.2,1 9.6]
Poland	47	[41.2, 52.9]	30.5	[29.4, 31.7]	16.5	[10.6, 22.4]
Estonia	49.4	[47.3, 51.5]	47	[44.5, 49.5]	2.4	[-0.8, 5.7]
N	30 494		41 753		72 247	

Source: Authors' analysis of SHARE and EHIS data. Estimates based on average marginal effects after a logistic regression model, setting age and sex to the all-country/all-survey means.

When we look at disability and employment, the picture changes slightly. Differences in the *absolute* level of disability employment (shown in Appendix B1) are relatively small – differences between surveys are still close to significance (p<0.06), despite the smaller sample size when looking only at people with disability, but the correlation between disability employment rates in EHIS and SHARE is very high (r=0.96). This is not always the case; a separate study comparing EU-SILC (similar in methodology to EHIS) to the European Social Survey (similar to SHARE), found greater inter-survey differences in the disability employment rate (country-level r=0.72) (Geiger *et al.*, 2017:Table A1). The cross-survey similarity for absolute disability employment in our case is therefore likely to owe more to luck than the inherent robustness of this comparison across surveys.

If we focus on the disability employment *gap* (Table 2.2), though, then we find considerable inter-survey differences. Firstly, the disability employment gap is generally higher in SHARE than in EHIS, possibly because of the different employment measure used in EHIS.[6] Secondly, looking at individual countries, there are a few large inter-survey differences in the disability employment gap. Sometimes the disability employment gap is smaller in countries and surveys that showed much greater disability prevalence (Denmark, Greece), which is what we would expect if people with minor limitations were more likely to report a disability in these countries (see also the following section and Geiger *et al.*, 2017:4). Surprisingly, though, sometimes the gap shows less easily explicable patterns (e.g., Poland, Czechia).

Table 2.2. Survey differences in the resulting disability employment gap in 50-69-year-olds

	SHARE	EHIS	Difference

Slovenia	14.6	[19.3 ,9.8]	11.8	[16.3 ,7.4]	-2.8	[-9.3, 3.7]
Italy	17.1	[22.7, 11.6]	13.0	[16.5, 9.6]	-4.1	[-10.6, 2.4]
Greece	18.0	[23.3, 12.6]	5.6	[10.4, 0.8]	-12.4	[-19.5, -5.2]
France	22.7	[27.9, 17.6]	20.5	[24.2, 16.7]	-2.3	[-8.7, 4.1]
Sweden	23.4	[28.4, 18.5]	14.4	[23.1, 5.7]	-9.1	[-19.1, 1.0]
Germany	23.9	[28.1, 19.7]	31.0	[35.0, 27.0]	7.1	[1.3, 12.9]
Austria	25.0	[30.0, 20.0]	19.8	[23.1, 16.5]	-5.2	[-11.2, 0.8]
Estonia	30.3	[35.2, 25.4]	35.8	[42.0, 29.5]	5.5	[-2.5, 13.4]
Denmark	30.4	[35.4, 25.5]	22.6	[28.5, 16.8]	-7.8	[-15.5, -0.1]
Poland	34.9	[47.7, 22.1]	22.1	[25.0, 19.3]	-12.8	[-25.9, 0.3]
Czechia	35.4	[42.8, 28.1]	26.3	[31.6, 21.1]	-9.1	[-18.1, -0.0]
N	30 494		41 753		72 247	

Source: Authors' analysis of SHARE and EHIS data. Estimates based on average marginal effects after a logistic regression model, setting age and sex to the all-country/all-survey means.

Overall, this example shows that when using more- vs. less-comparable surveys, we will sometimes see similar results (here, for absolute disability employment rates), but sometimes find quite different results (here, for disability prevalence and disability employment gaps). The wider literature further emphasises this point, with different surveys showing different cross-country patterns and different within-country trends (Baumberg *et al.*, 2015; Berger *et al.*, 2015; Croezen *et al.*, 2013; Geiger *et al.*, 2017; Rubio-Valverde *et al.*, 2019). If we use a less-comparable survey, we can never be sure that the results are the same as we would find if we made greater attempts to reach comparability. Whatever the data that we can use, it is crucial to consider the possibility that apparent differences between countries in fact reflect methodological incomparability rather than real-world variations.

Accounting for the varying prevalence of disability

A further way of trying to deal with issues of limited cross-national comparability is to use a prevalence-adjusted measure, which is simply the conventional disability employment gap multiplied by the reported prevalence of disability (i.e., $gap \times prevalence$). This was first proposed by Richard Berthoud (2011) but the measure is also used in other works, including the most recent OECD report on disability policy trends (Jones and Wass, 2013; OECD, 2022). We follow Jones and Wass in terming it the 'prevalence-adjusted disability employment gap'. One intuitive way of thinking about this measure is that – if the employment gap measure shows the *causal* impact of disability on employment – then the prevalence-adjusted measure shows the total share of the population prevented from working due to disability.[7]

While the prevalence-adjusted gap has previously been used, its value and limitations have not previously been clearly explained – so this is our focus for the rest of the chapter. We also calculate confidence intervals around the prevalence-adjusted gap, which again has not been done before.[8]

The advantage of the prevalence-adjusted measure is that it can cope with situations in which people with minor limitations sometimes do and sometimes do not report a 'disability'. To understand this, it helps to imagine a group of people with limitations that do not affect their chances of working at all which is not implausible (DWP and DH, 2016 Chart 1.6; Jones, 2006). The conventional disability employment gap is very sensitive to this group: the more of them who report a disability, the smaller the employment gap (because it includes people with no employment penalty). In contrast, the prevalence-adjusted gap is completely unbiased no matter how many of this group report a disability.

In practice, things will not be this simple. Box 2.1 explains in more detail the bias in the conventional disability employment gap and the prevalence-adjusted gap in different scenarios. The takeaway message is that the conventional gap and the prevalence-adjusted gap are useful complements to each other –

where these measures show different patterns, it alerts us to the possibility that differences across time or place might be because of reporting differences. It also directs our attention to the prevalence of disability, which needs to be understood if we are to obtain a complete picture about the extent of disability inclusion and exclusion in the labour market. Alongside other measures, it seems sensible that policymakers should know the percentage of the population who are potentially prevented from working due to disability.

Box 2.1. How far does the prevalence-adjusted gap measure reduce bias?

In this report, we suggest using an adjusted measure to look at disability and employment – the 'prevalence-adjusted disability employment gap'.

The advantages of this measure have not previously been set out. The easiest way of explaining the advantages of this measure is to imagine three groups of people:

- Always-disabled people (who report a disability in all times/places),
- Never-disabled people (who never report a disability), and
- Sometimes-disabled people (who only sometimes report a disability).

To compare biases in different measures, we then look at how the disability employment gap changes if sometimes-disabled people all report a disability, vs. when none of them report a disability. Note that nothing changes substantively; the only thing that changes is the *reporting* of disability.

Bias in the simplest scenario

In the simplest situation, we assume that the employment rate of sometimes-disabled people ('ESD') is the same as the rate for never-disabled people (ESD=ND). This is a plausible assumption, but we return below to consider how biased measures would be if this is not the case.

The conventional disability employment gap

Remember that the disability employment gap should not change, because no-one's employment situation is changing – it is simply the reporting behaviour of sometimes-disabled people that has changed. But using some simple algebra (which is given in Appendix C), we can see that the conventional disability employment gap does change, as follows:

$$\text{Change in gap as reporting changes} = (E_{ND} - E_D)\frac{P_{SD}}{(P_D + P_{SD})}$$

That is**, the change depends on the difference between the employment rates of never-disabled and always-disabled people ($E_{ND} - E_D$), multiplied by how many sometimes-disabled people there are compared to always-disabled people ($\frac{P_{SD}}{(P_D+P_{SD})}$). This could be quite big: under realistic scenarios, the disability** employment gap could rise from 30% to 40% simply due to reporting changes.

The prevalence-adjusted disability employment gap

The prevalence-adjusted disability employment gap is simply the prevalence of disability multiplied by the disability employment gap. Despite its simplicity, though, it helps us deal with this bias.

Again, using simple algebra (given in Appendix C), if we look at this hypothetical scenario – where all the sometimes-disabled people change from reporting no disability to reporting a disability – then the change in the prevalence-adjusted gap is precisely zero. That is, the prevalence-adjusted gap is fully robust to the bias caused by reporting changes, if the employment rate for sometimes-disabled people is the same as the employment rate for people without disability.

Bias in other scenarios

In practice, it may well be the case that the employment rate for sometimes-disabled people is slightly lower than that of never-disabled people. The formula for the bias of the conventional disability employment gap does not take a simple form here (see Appendix C). Nevertheless, by plugging-in plausible values, we can see that:

- There are some situations where the disability employment gap does not change, while the prevalence-adjusted gap rises substantially (from 5.7% to 7.6%). In this case, it is when the employment rate of sometimes-disabled people is 47.5% - that is, only slightly higher than the employment rate for always-disabled people (40%).
- Where the employment rate for sometimes-disabled people with disability is between 47.5% and 80% (the employment rate for people without disability), the disability employment gap falls, while the prevalence-adjusted gap rises. The size of these changes depends on exactly what the sometimes-disabled employment rate is – the closer it is to 80%, the greater the change in the disability employment gap, and the smaller the change in prevalence-adjusted gap.

In summary, this clarifies that there are three reasons to look at the prevalence-adjusted gap in addition to the disability employment gap, as done in OECD (2022:38-43):

1. Firstly, when sometimes-disabled people have an employment rate that is the same as never-disabled people, the prevalence-adjusted gap is unbiased (even though the disability employment gap is biased).
2. Secondly, when sometimes-disabled people have an employment rate that is only slightly lower than never-disabled people – which in our view is the most likely scenario –, the prevalence-adjusted gap is much less biased than the disability employment gap.
3. Third, when sometimes-disabled people have an employment rate that is much lower than for never-disabled people, then the disability employment gap will fall, while prevalence-adjusted gap will rise. Where there are different trends in the two measures, this therefore tells us that changes over time might be due to reporting changes.

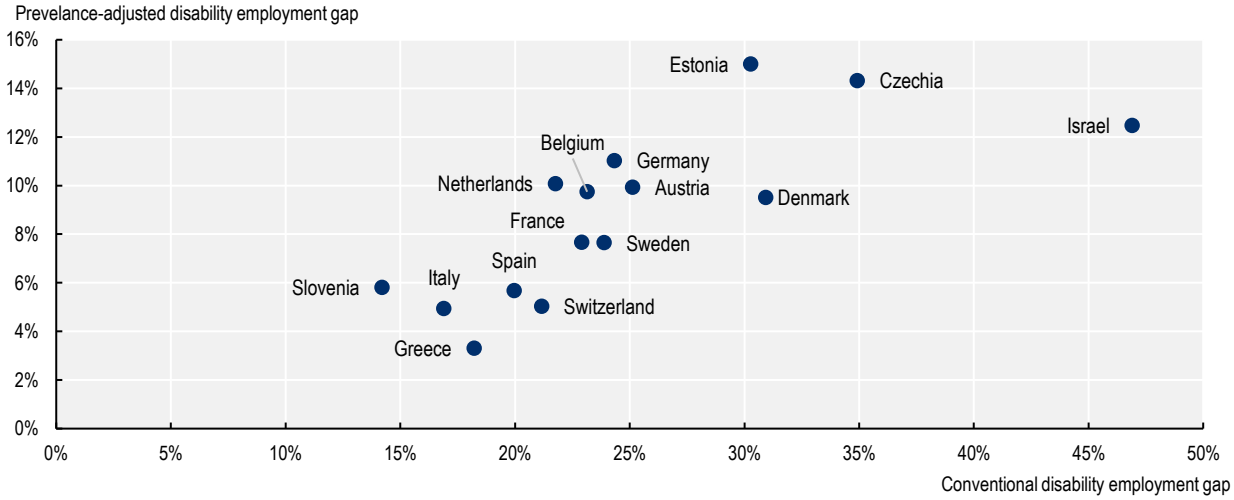
Two case studies showing the impact of the prevalence-adjusted gap measure

The prevalence-adjusted gap vs. the conventional gap: international comparisons

To illustrate the use of the prevalence-adjusted gap and how it compares to the conventional disability employment gap, we use the SHARE data used in the previous section but now for all available countries, rather than only those that overlap with EHIS countries. The results are shown in Figure 2.1.

Overall, the picture is similar using the two different measures (they correlate quite strongly, r=0.68). The main exception is for Israel, which has the highest conventional disability employment gap by some distance, but only the 3rd-highest prevalence-adjusted gap – or, put another way, the worst disadvantage experienced by people with disability, but only the third-worst proportion of people potentially prevented from working due to disability. This is because the self-reported prevalence of disability in Israel is relatively low; it is plausible that this is partly because fewer people with minor limitations report a disability there, relative to other countries, which skews the disability employment gap measure. Similarly, Slovenia has the lowest disability employment gap, but the 5th-lowest prevalence-adjusted gap. While there is a strong similarity in the comparative picture we obtain from the different measures, there are also some striking differences.

Figure 2.1. The conventional gap vs. the prevalence-adjusted gap, for 50-69-year-olds, 2010-14



Source: authors' analysis of SHARE data. Estimates based on average marginal effects after a logistic regression model, setting age and sex to the all-country/all-survey means. Exact figures and confidence intervals are given in Appendix B.

The prevalence-adjusted gap vs. the conventional gap: the United Kingdom case

Another useful way of understanding the value of the prevalence-adjusted gap is to look at the case of the United Kingdom, where the conventional disability employment gap in the past has been an important measure for policy, and for discussion. For example, halving the gap for the UK working-age population was the headline aim of a 2017 Government strategy consultation. This was turned into the aim to get *'a million more people with disability into work'* in the final strategy, using the same data as for the disability employment gap (i.e., the UK Labour Force Survey, or 'LFS').[9]

These measures both seem to show an unqualified success story. The goal of an extra million people with disability in work in the United Kingdom was reached in 2022,[10] while the conventional disability employment gap showed a steep fall between 1998 and 2023, from 42 to 29 percentage points (see Appendix D3 for methodological details). Yet, many observers (including academics and Parliamentary scrutiny committees) have been sceptical that this reflects improved disability inclusion.[11] Astonishingly, this is a period in which the prevalence of single-item activity-limiting disability (using a measure similar to the GALI) rose by 50%, from 15.4 to 23.0% of the population, with most of the rises happening in the years 1998-2001 and (in particular) a sharp and steady rise in the period 2018-2022.

There are strong reasons to be sceptical that this sharp rise in disability prevalence represents a genuine increase in impairments and activity limitations. Partly this is because the earlier trend (1998-2001) is not robust across different surveys; two other major government surveys show stable trends in disability (Baumberg *et al.*, 2015). Even more convincingly, the most careful efforts to summarise trends in ill-health and disability across different domains show no rises in disability in the earlier part of this period. This includes the best international collaborative effort to understand trends in disability – the Global Burden of Disease study – that shows effectively no change in disability in the United Kingdom in 1998 to 2019;[12] and a systematic review of every comparable trend series from the Health Survey for England in the period 1994-2014 (Geiger, 2020). The 2018-22 rise is too recent to have been subject to the same scrutiny, but given the earlier evidence, it is clearly possible if not likely that it reflects reporting changes rather than changes in impairments and activity limitations.

It therefore seems likely that this single-item activity-limiting disability measure in the LFS is not comparable over time, at least not for the United Kingdom, and that the conventional disability employment gap is,

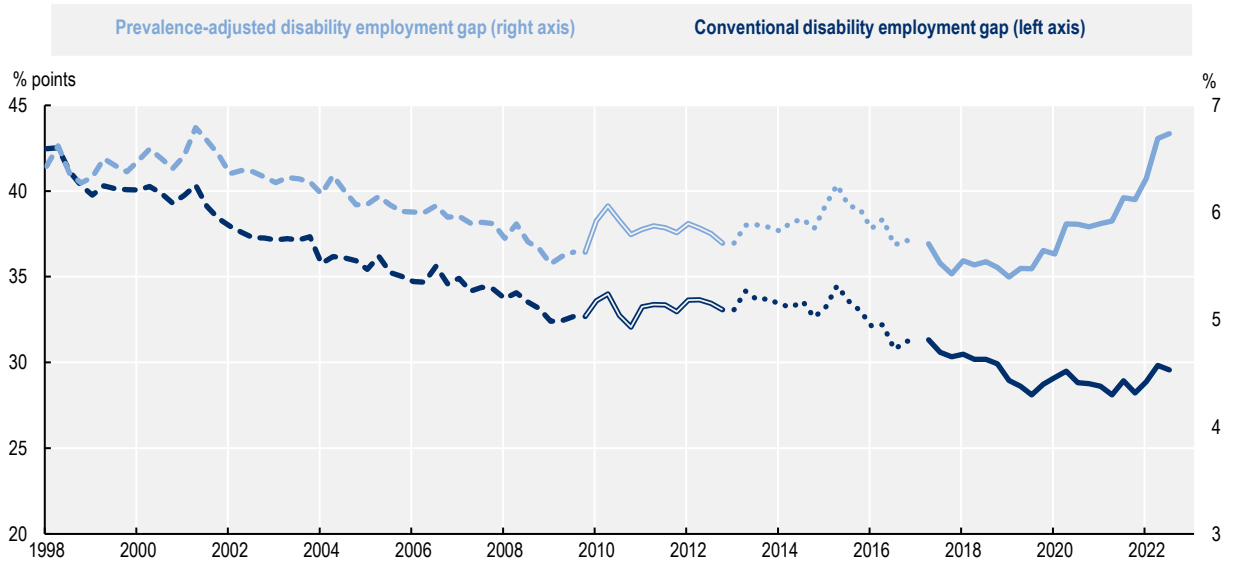
therefore, also misleading. In recent publications, analysts from the Department of Work and Pensions have drawn attention to these limitations – for example, in saying that the disability employment gap is affected by the rate of disability *per se*, or even explicitly estimating how far changing disability prevalence explains the increasing number of people with disability in employment.[13] But no alternative headline measure has been used or proposed in place of the conventional disability employment gap, to which politicians have continued to refer.[14]

How does the prevalence-adjusted gap measure cope with rising (reported) prevalence? The results are shown in Figure 2.2. The conventional disability employment gap shows the steady improvement noted above. However, using the prevalence-adjusted gap, we see a mixed picture – an improvement in the period 2002-2010, little change in 2010-2020, and a sharp rise since 2020. At the end of this period, the proportion of people potentially prevented from working due to disability (measured by the prevalence-adjusted gap) is even worse than the situation at the start of the period in 1998.

We cannot be certain that the prevalence-adjusted gap is correct here – we are still relying on a problematic disability measure, and the bias in the prevalence-adjusted gap depends on the employment rate of people who would not have reported a disability in 1998 but do now, something which we simply do not know (see Box 2.1). Nor do we have a better measure to compare it to using the techniques in Chapter 3 (although the results in Chapter 4 show a similar picture to the prevalence-adjusted employfment gap).

What we can say with certainty is that it is clearly misleading to rely solely on the conventional disability employment gap in a time of sharply rising (or, changing) reporting of disability prevalence. By using the prevalence-adjusted gap measure alongside the conventional gap, this alerts us that changes in disability reporting might be playing a role. Here the prevalence-adjusted gap provides an easily available alternative measure under different (and often more plausible) assumptions.

Figure 2.2. Comparing two measures of disability employment gaps in the United Kingdom



Source: official statistics for the working-age population (aged 16-59(f)/65(m) until 2009, 16-64 from 2010-) from the Labour Force Survey, using a chained series to account for discontinuities in the data (shown by different line patterns above); see Appendix D3.

Conclusions on the prevalence-adjusted employment gap

Overall, the prevalence-adjusted gap is a useful complement to the disability employment gap. This is partly to ensure that policymakers pay attention to the share of the population that is potentially prevented

from working due to disability (rather than just the disadvantage associated with disability). But partly it is because it can highlight comparisons that are likely to reflect methodological artefacts rather than genuine differences in disability inclusion or exclusion.

The value of using the prevalence-adjusted gap in this way can be seen in the findings above. While there are some situations where the prevalence-adjusted gap produces findings like the conventional disability employment gap (e.g., comparing 50–69-year-olds across most countries), at other times the results are quite different (e.g., for Israel in comparative perspective, and for trends over time in the United Kingdom). This mirrors our findings in the first part of the chapter when comparing survey comparability.[15]

The takeaway message is that the conventional gap and the prevalence-adjusted gap are useful complements to each other – they highlight different things that policymakers need to know; and where they show different patterns, they alert us to the possibility that we should not naïvely trust single-item activity-limiting disability measures. For more robust comparisons, though, we need to go beyond these measures, using data and techniques that we explore in the following chapter.

3 Measuring the disability employment gap by combining multiple impairment and activity limitation indicators

More robust comparisons require better measures of disability. The best way of doing this is to use a series of specific questions on impairments and activity limitations, which are answered reasonably similarly wherever and whenever they are asked, which are then combined into a single scale. The resulting disability measure refers to people with impairments who are *potentially disadvantaged* in the environment in which they live, which in turn enables us to see how policies affect disability inclusion.

There is some confusion about how best to implement this. We explain what is at stake in these choices, and recommend that:

(1) The items included should be comparable and comprehensive.

(2) The weights for each item in the scale should be produced using the 'predicted disability' approach – that is, by weighting each item according to how strongly it predicts self-reported disability.

(3) The scale should be turned into a binary disability category using a probabilistic approach, a new approach we explain in detail.

Self-reported single-item measures of activity-limiting disability are likely to be answered in different ways by different people in different times and different places. They are particularly problematic because they are influenced by disability-related policies themselves. Such measures are therefore an imperfect way to compare countries and examine trends, and especially to evaluate policies.

The best solution to the limited comparability of self-reported activity-limiting disability is to use questions that are likely to be answered similarly in different contexts – questions on impairments and on activity limitations – and then combine these into a summary measure. This chapter explains how to follow such an approach, splitting it into a series of different decisions:

- The logic of the overall approach.
- Choosing the multiple measures to include in the scale.
- Combining these measures into a single scale.
- Whether to turn this scale into a binary indicator, and if so, how.

It is important to be aware of the disagreements about disability measurement between different groups (as seen, e.g., in a high-profile debate about disability measures in the United States).[16] This chapter draws on the insights from multiple different groups, including the Washington Group on Disability Statistics, the National Bureau of Economic Research, and the World Health Organization (the World Disability Report, the WHO Disability Assessment Schedule and the WHO Model Disability Survey), recognising that all of these approaches are trying to work towards similar goals, and there is something of value in all of them.

The logic of focusing on impairments and activity limitations

The key issue, as we have said before, is that we cannot have a disability measure that itself already factors in the role of policies and the wider environment, because this makes it impossible to tease apart how policies affect the inclusion of people with disability. As we explain in more detail just below, we are still understanding disability as the interaction between impairments and environments (and more broadly, disability assessments for benefits and services certainly need to take the environment into account) – but the challenge is how to produce robust measures within this approach.

Our view[17] is that a robust measure should be based on a series of specific questions on impairments and activity limitations, which are answered reasonably similarly wherever and whenever they are asked, which are then combined into a single scale. To make this more concrete, let us use an example of an activity limitations question, *'raising a 2-liter bottle of water or soda from waist to eye level'*. People's answers to this question are unlikely to be affected by the nature of paid work or the benefits system in their country. In contrast, policies and work environments may well affect whether someone reports a *'health problem or disability that limits the kind or amount of paid work you can do'*. The most robust measures, in our view, combine a series of questions on specific aspects of impairments and activity limitations. (An alternative view is that we do not need to worry about which questions we ask as we can account for any biases statistically; this is discussed in Box 3.1).

While we believe that questions on activity limitations are more comparable across different environments, it is difficult to empirically demonstrate this in the absence of an agreed 'gold standard' to benchmark against. The best we can do is to note that there are implausibly large cross-national differences in activity-limiting disability, but smaller, more plausible differences in specific measures (Croezen *et al.*, 2013). Moreover, on a conceptual level, we must accept that even activity limitation measures are not perfectly comparable.[18] Some cultural differences may remain; as Mont (2007) notes, *"'dressing oneself' can take on very different connotations in a society where one ordinarily slips into pants and a loose fitting shirt, compared to dressing in something as complicated as a sari."* These issues are magnified when considering the impairment questions often used to capture mental health, pain and energy limitations, which are inherently subjective and will be answered differently across time and place. For example, even

the well-validated depression scale used in SHARE does not fully show measurement invariance across countries (Castro-Costa *et al.*, 2008; Maskileyson *et al.*, 2021).

Despite these caveats, though, symptom or feeling-based measures of mental health and pain – particularly with well-validated symptom scales – are *prima facie* likely to be much more comparable than reporting of medical labels or general assessments of participation restrictions. And more broadly, it is still reasonable to assume that more specific impairment and activity limitation measures are interpreted more consistently across different environments than general activity-limiting disability measures like the GALI. In the next section we discuss exactly which activity limitation and impairment measures need to be included in the scale. Before this, we discuss whether this approach is compatible with the ICF.

Is defining disability through impairment and activity limitation compatible with the ICF?

A possible objection here is that a focus on impairments and activity limitations seems incompatible with the ICF (and indeed, with the social model of disability).[19] This is not the case, either for our approach, or for others making similar calls to focus disability measurement on impairment and activity limitations (Madans *et al.*, 2011; Mont, 2019). But it is important to explain exactly why this objection is wrong.

While the ICF and the social model of disability differ in many ways, both models understand that participation restrictions – including work disability – do not flow straightforwardly from impairments. Rather, impairments only result in work disability in disabling social environments, where some people are excluded from full participation because of the way that work and wider society is arranged (Mont, 2019). Within the ICF, this is sometimes referred to as the difference between 'intrinsic health capacity' and real-world 'performance', as we described above.

In this context, 'disability' can either mean (i) people with impairments who are *actually disadvantaged* given their particular environment ('performance'); or (ii) people with impairments who are *potentially disadvantaged* in their environments ('capacity'), even if they are in an environment in which they do not actually face any disadvantages themselves. While policymakers are mostly interested in actual disadvantage, we cannot base a robust measure of disability on this. Partly this is for measures based on participation restrictions, we cannot tell if a reduced level of disability means that impairments have become less common, or whether social environments have become less disabling. But it is primarily because it is impossible to measure participation restrictions in a way that is unaffected by policies themselves (see Chapter 1) – if you create a less disabling social environment, then this will reduce the number of people reporting participation restrictions. Such a policy could, therefore, appear to be a failure because the barriers of people who still report participation restrictions may be more severe. If we measure participation restrictions directly, it is impossible to track how the policies have affected the lives of people with disability (rather than how policies affect whether people *report* a disability). [20]

Instead, when doing international comparisons using a scale based on capacity, we maintain the ICF/social model approach in the way we *interpret* the data (Mont, 2019). As Palmer & Harley (2011) put it, *"while disability measurement questions address only limited aspects of the ICF model, the analysis and interpretation can incorporate the model more fully."* A scale based on impairments and activity limitations enables us to identify people who are potentially work-disabled in OECD contexts, and to do this in a consistent way that will not be affected by policy-induced changes in how people report disability. The extent to which this group are excluded from a given society – e.g., in terms of employment rates or poverty levels – is then a measure of how disabling that society is. Rather than measuring participation restrictions directly, it is better to measure disability in a robust way using impairments and activity limitations, and then to use the resulting disability employment gap as a robust measure of participation restrictions.

We are committed to the core insights of the ICF, and furthermore, that disability assessments within social protection systems should be based on performance rather than capacity as we have argued elsewhere (Geiger *et al.*, 2018). But to empirically compare how disabling social environments are, or how effective

policies are, we need an underlying measure of disability that is measured consistently in different times and places, rather than a measure that means different things in different social environments.

Step #1: Choosing impairment and activity limitations measures for the scale

We have argued that the best way of measuring disability consistently is to use a set of specific questions on particular aspects of disability which are answered similarly wherever and whenever they are asked, and then to combine them. This raises the question about which specific measures to use.

In previous studies that use scales, there is often little discussion about this – it is simply a matter of putting in all the available measures. Such an approach means that questions that are unlikely to be interpreted consistently over time and place are included. For example, one prominent scale developed by the US National Bureau of Economic Research (Poterba *et al.*, 2013; Wise, 2017) includes general self-reported health, which we have already seen has comparability issues. It is also common to use medically diagnosed chronic health conditions (Jürges, 2007; Poterba *et al.*, 2013), which we know vary considerably over time and place because of changes in medical knowledge and practice.[21]

More robust analyses require us to think more carefully about which measures to use. In an ideal world, we would use…

1. …measures that are likely to be comparable across time/place, which means excluding measures based on general health, participation restrictions or medical labels *(comparability)*; and
2. …measures that cover all dimensions of impairments/activity limitations that are likely to be work-disabling in OECD countries, including mental health, pain, and energy limitations *(coverage)*.

There is an inherent trade-off here: some of the things that are important to cover are also difficult to measure comparably, such as mental health. There is no perfect answer here, and different researchers will make different decisions. Moreover, coverage also leads to a very long survey instrument, particularly compared to the single-item GALI measure of activity limiting disability. In Chapter 5, we recommend an efficient way of balancing these cost and comparability considerations in an optimal way, while emphasising that a high priority for future research is to invest in a dataset that allows us to robustly measure and compare disability prevalence across time and place.

Several international and national groups have spent considerable time and effort developing comprehensive sets of questions around disability, as shown in Box 3.1.

Box 3.1. Question sets designed to measure disability in the general population

The WHO Model Disability Survey (MDS) was developed by the WHO and World Bank to fill the gap in ICF-based question sets to monitor population disability (WHO and World Bank, 2011). The full version of MDS is extremely long (90 questions), but shorter versions have been proposed, including a 24-item 'brief MDS' (Sabariego et al., 2022), and the 11- item 'FDD11' (Lee et al., 2022). The MDS is closely associated with a particular statistical approach that uses latent variable modelling (Sabariego et al., 2022; Sabariego et al., 2021), which is discussed further below. The major strength of MDS is a rigorously designed set of questions on impairments and activity limitations across key domains, which has undergone more extensive validity testing than most other scales (linked to this type of latent variable modelling). However, it also includes questions on participation restrictions that are not likely to be comparable across different environments (e.g., the performance module asks one question about challenges in getting things done at work (or at school), while the capacity module includes a question on self-reported general health) and would therefore need to be excluded when using the approach recommended in this report.

The Washington Group have developed a set of activity limitation measures that fit within the impairments/activity limitations-focused approach outlined above. We do not focus on the short 'core set' of six questions, which misses a number of domains of functioning (Sabariego et al., 2015). There is however better coverage from the 'enhanced core set' that further includes four questions on anxiety/depression, and the 'extended set' that covers 10 dimensions via 34-37 questions. While the core set does not capture pain, energy-limiting impairments[22] or mental health (Molden and Tøssebro, 2010; Mont, 2007), these are partly covered in the longer question sets, albeit using a different approach from questions in the rest of the survey.[23] The scales have been tested in several ways (Altman, 2016; Miller et al., 2011), but there have been arguments about the exact wording of the questions and response categories,[24] as well as how the scales are turned into binary disability measures (see below).

The OECD's International Survey of People Living with Chronic Conditions: within the framework of the Patient-Reported Indicator Surveys initiative (PaRIS), the OECD has launched the largest international survey to date, focusing on primary care service users aged 45+, covering 19 countries and more than 100 000 patients. While it focuses on aspects that matter most for patients, it includes several relevant question sets adapted from elsewhere, brought together into the PROMIS-10 questionnaire, to which it adds measures on mental health. PROMIS-10 is internationally validated (Fischer *et al.*, 2018; Plessen *et al.*, 2024; Terwee *et al.*, 2021; Terwee and Roorda, 2023), and PaRIS supplements this with field trials and cognitive tests across participating countries. While the PaRIS survey does not cover the whole population, it provides another carefully constructed and validated question set that could be used for international comparisons. For more details see http://www.oecd.org/health/PaRIS

There are also many other notable sets of questions, including *inter alia* the WHO Disability Assessment Schedule (which is similar to WHO MDS in approach but designed to be used in disability assessments; Bickenbach, 2011; Saltychev *et al.*, 2021; Üstün *et al.*, 2010), the WHO World Health Survey (as used in the World Disability Report), and the Canadian *Disability Screening Questions* ('DSQ').[25] All of these question sets are a substantial achievement, requiring much investment of time, resource and expertise, and many of these have been used in numerous studies globally (as reviewed by Casebolt, 2021; Federici *et al.*, 2017; Groce and Mont, 2017; Loeb, 2016; Sabariego *et al.*, 2022). This has come alongside fierce debate over which of these question sets should be preferred. However, none of these can be used in this report to study the disability employment gap – simply because none of these question sets are included in any comparable cross-national surveys covering the general population in OECD countries. Whichever question set is used, the priority is to ensure that we have at least *some* detailed, comparable data on specific measures of impairments/activity limitations in the future.

In the rest of this chapter, to illustrate our approach, we use SHARE data which has the largest number of available impairment and activity limitation measures of any input-harmonised survey. This includes activities of daily living (ADLs), instrumental activities of daily living (IADLs), motor skills impairments, vision and hearing limitations, and a validated mental health symptom scale (see below). We should note, however, that some of these measures used in SHARE may be interpreted differently in different countries, that this is not an exhaustive list of potentially work-relevant impairments/activity limitations, and that the survey only covers people aged 50 years and over. Still, this set of measures manages to cover most of the major domains relevant to work limitations, and more than any other input-harmonized survey of any part of the working-age population that we know of.

However, a key limitation of SHARE is that this survey covers the population aged 50 years and over only. In Chapter 4, therefore, we also use similar techniques on the smaller list of impairment and activity limitation measures available in EHIS data, which cover all age groups.

Step #2: How to create a disability scale from multiple measures

When using a set of impairment and activity limitation measures, the next issue is how to combine multiple measures into a single summary disability measure. While there are many examples of creating disability scales, researchers often seem unaware of the limitations of their methods, and how different choices might produce more robust results.

The most common way that these multiple indicators are turned into a single scale is by using a latent variable approach, e.g., factor analysis or Item Response Theory (IRT). This underpins the WHO's proposals for using its latest surveys (WHODAS and MDS), the World Disability Report (WHO and World Bank, 2011) and the latest phase of the National Bureau of Economic Research (NBER)'s 12-country project on 'Social Security Programs and Retirement around the World' (Wise, 2017) – a series that has previously been described as *"hands down the most influential use of international comparisons in economics"* (Banks and Smith, 2012). These statistical techniques all weight different items in the disability scale based on how much they seem to be measuring the same thing (an unobserved, 'latent' variable), which we hereafter term the *'latent disability' approach*.

However, these weights – which are based on the associations between the measures – are not sensible. Imagine if we have nine measures of motor skill impairments, and a single measure of mental ill-health (which is close to the situation in the NBER data). The nine motor skill measures will be closely correlated with each other, but less strongly to the mental ill-health measure. As a result, the mental ill-health measure will receive a low weight in the scale, which is problematic in two ways. It makes no conceptual sense for mental ill-health to have a low weight, given how important it is for disability in the real world. And it makes no methodological sense, as the disability weights will be strongly affected by which measures are included in constructing the scale and are, therefore, very inconsistent across applications.

A better approach is to weight the impairment and activity limitation measures by how strongly they predict participation restrictions (what we term the *'predicted disability' approach*). This is similar to the comparative research by Jürges (2007), though the idea of constructing an 'objective' health index in this way goes back at least to Bound (1991). Jürges in his work regressed a series of different specific health measures on self-reported general health, finding, e.g., that cataracts (the least disabling health condition) lowered health by only 0.077, while Parkinson's disease (the most disabling condition) lowered it by ten times as much (0.859). These are then used as weights in a health index, so, e.g., someone with cataracts would have a disability score of 0.077, while someone with Parkinson's disease would have a score of 0.859. (This is a slight simplification; see below for further details). Our approach is similar, but we focus on how well these different impairment and activity limitation measures predict *self-reported activity-limiting disability*, rather than self-reported general health.

It is sometimes claimed that the latent disability approach can statistically solve problems of comparability across time and place, which would be a reason to use this approach even if the weights themselves are not sensible. We explain this claim in Box 3.1, where we further explain why we think this claim is more complex than it appears – this is an area where further discussions would be welcome to broker common understanding and potentially agreement between different researchers and approaches.

Illustrating different methods for constructing disability scales

To illustrate the difference that the various choices make, in the following example we compare the disability employment gap across three measures of disability: (i) self-reported activity-limiting disability (as in Chapter 2), (ii) latent disability, and (iii) predicted disability.

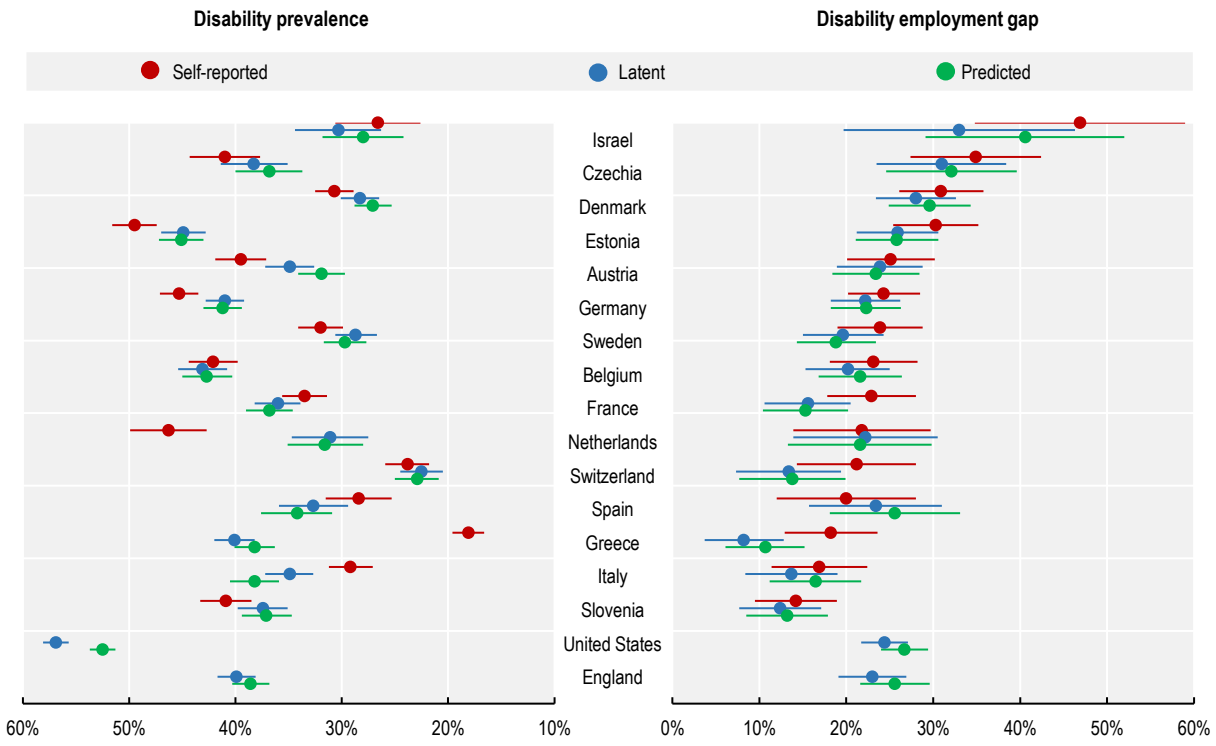
The best-quality international comparative data on health and disability are the Global Aging Surveys, here including SHARE (for much of Europe) as well as the matching surveys for England (English Longitudinal Study of Ageing; ELSA) and the United States (Health and Retirement Study; HRS).[26] Unlike for the analysis in Chapter 2, our comparisons can now also include England and the United States.[27] We again use the 2010-14 data, but now also use a series of individual health measures: six measures of ADLs; another six measures of IADLs; ten motor skill impairments; single-item summary measures of each of vision limitations and hearing limitations; and a mental health scale – full details are given in Appendix D1. The mental health scales are slightly different between surveys, though this is unlikely to affect our results.[28] While this list partially covers mental health-related limitations (including in the IADLs), it contains more items measuring physical health-related limitations, which are therefore likely to receive higher weights in the latent disability scale.

To summarise the technical details: latent disability is created using hybrid Item Response Theory (IRT) models, which are better able to cope with our categorical health measures than the techniques used in some of the wider literature.[29] Predicted disability is created using a straightforward regression model: we regress the multiple health/impairment measures on single-item activity-limiting disability using a logistic model. To try to ensure that the weights reflect real associations between these measures and disability, rather than socioeconomic patterning or differences in reporting styles, the logistic model also controls for country (similar to Jürges, 2007), employment status (to account for work-related biases in reporting disability), age (single year), gender, and education. We then use this model to estimate the predicted probability of each respondent reporting a disability, based purely on these limitation measures (holding the control variables constant). Readers who want to replicate this approach can find further detail in Appendix A, including a sample code.[30]

To enable comparisons to the binary GALI-style measure in Chapter 2, we turn the latent and predicted disability scales into binary measures of disability. For the purposes of this illustration, this is based on a cut-off on each scale that ensures the resulting binary variable matches the self-reported prevalence of disability (this decision and further alternatives are discussed in the next sub-section).

Our results are shown in Figure 3.1 (exact figures are given in Appendix B; we return in Chapter 4 to considering the absolute disability employment rate alongside the disability employment gap). Looking first at the prevalence of disability (left-hand panel), we find that self-reported activity-limiting disability is highest in the Netherlands, Germany and Estonia, and lowest in Mediterranean countries (Spain, Italy, Israel and particularly Greece) plus Switzerland. When we use the disability scales rather than single-item activity-limiting disability, though, the extent of differences between countries reduces sharply: countries with the highest disability levels fall closer to the average (e.g., the Netherlands), while some of those with lower disability levels have higher levels (e.g., Greece).

Figure 3.1. How international comparisons change when using single-item activity-limiting vs. latent or predicted disability scales, for 50-69-year-olds, 2010-14



Source: Authors' analysis of SHARE-ELSA-HRS data. Estimates based on average marginal effects after logistic regression models, setting age and sex to the all-country/all-survey means. Exact figures and confidence intervals are in Appendix B.

If we turn to the disability employment gap, we find the picture is much more similar between self-reported measures and multi-item scale measures of disability. There are some differences – e.g., the gap in the Netherlands goes up, the gap in Switzerland goes down – but even in these cases differences are small relative to the confidence intervals, and across most countries the two measures look similar. (This may seem surprising, given the prevalence of disability is so different in some countries. We work through an example in Appendix B3, but the main message is that when changing disability measure, the extent of changes in the disability employment gap depends primarily on the employment rates among groups with changing disability, not the changing disability prevalence.) It is therefore possible that large changes between measures in the *prevalence* of disability have little effect on the *disability employment gap*.

However, there is no logical requirement that these measures produce similar disability employment gaps. We show this clearly in Chapter 4 below using EHIS data, where we find very different results for predicted (vs. self-reported) disability. Just as in the previous chapter, these measures sometimes produce similar results and sometimes produce an entirely different picture. We consider the methodological and substantive implications of this in Chapter 5.

Within the two multi-item disability scales, there are also times when latent and predicted disability give different pictures – including the disability employment gaps in the United States, France and Greece. In nearly all cases the disability employment gap for predicted disability is higher than it is for latent disability, likely because the disability weights better capture the genuine barriers people face. However, there are fewer differences between predicted disability and latent disability in terms of prevalence – only in Estonia and Czechia (and to a lesser extent the United States and England) is there a noticeable difference.

We explore the reasons for the similarity between the latent and predicted scales in Appendix B4. In short: the different multi-item scales produce different disability weights, but similar binary disability variables (at least in this case) – and therefore produce similar comparisons across countries.

Step #3: How to turn the index into a binary measure of disability

We have argued that it is better to measure disability using impairment and activity limitation-based scales (rather than short activity-limiting disability measures), and that the best way of creating these scales is to use predicted (rather than latent) disability. There is one final important issue to consider: should we turn these *continuous* disability scales into a *binary* measure of disability? And if so, how should we do this?

In terms of whether we *should* use binary measures, there is clearly no easy dividing line between people with vs. without disability. Instead, there are a larger group of people who face at least some barriers, with progressively smaller groups facing progressively larger barriers (Burkhauser *et al.*, 2014:196). Given that we have created a continuous disability scale in the previous section, we could analyse this scale without turning this into a binary variable (e.g. Geiger *et al.*, 2019; Jürges, 2007; Poterba *et al.*, 2013). However, there are a number of disadvantages to using a continuous scale.[31] Instead, it is clearer to create multiple disability categories – e.g. mild, moderate and severe disability (Sabariego *et al.*, 2025) – and look at the disability employment gap for each one. We do this in our example below, splitting between those with a low, medium and high chance of reporting a disability.

Compared to those with low chances of reporting a disability, we find that disability employment gaps for people with medium chances of reporting a disability are only moderately correlated at a country level (r=0.48) with the gaps for those with high chances of reporting a disability (see Appendix B) – illustrating how policy conclusions that ignore the differences between these groups may be misleading.

In terms of *how* we create binary measures, in nearly all the previous research that we are aware of, researchers have used a fixed cut-off for their disability scale, with everyone below the cut-off counted as 'not disabled', and everyone above the cut-off counted as 'disabled'. The cut-off point *"is always the key issue in disability statistics"* (Shakespeare, 2013), and is always somewhat arbitrary, but there are various cut-offs that have been justified:

1. The most common cut-off for a scale has been set to whatever matches the all-country prevalence of single-item activity-limiting disability (as we do above) – so for example, if 23% of people self-reported a disability using a GALI-style measure, then our cut-off point in the scale is whatever value 23% of our sample is above. While arbitrary, this is politically sensible given that it can be problematic when new disability measures reduce the headline count of people with disability.
2. The WHO's preferred approach in some cases (e.g. in the World Disability Report 2011) is to look at the average disability score among (i) people reporting extreme limitations in any domain, and (ii) people reporting chronic conditions (arthritis, angina, asthma, diabetes, depression). However, the WHO MDS team use more arbitrary cut-offs, in order to divide their disability scale into four groups (none / mild / moderate / severe) (Sabariego *et al.*, 2022).
3. When looking at mental health, the OECD has sometimes used a fixed cut-off so that the highest 20% of scores in each country are classified as having 'mental ill-health'. This has been done to overcome the problem with different mental health scales available in different countries and over time (e.g. OECD, 2015:28); an approach backed up by epidemiological studies which find no large differences across countries in the prevalence of mental ill-health, at least prior to the COVID-19 pandemic. This approach is only meaningful if the focus is on social and economic outcomes of this group, as the prevalence is fixed. A similar approach could be used for disability scales, if we wanted to assume that the prevalence of disability was the same in all countries.[32]

4. The Washington Group method is potentially the simplest: anyone who reports *any* of a series of impairments and activity limitations beyond a given threshold (generally 'a lot of difficulty' in a domain) is classified as having a disability. This approach has been criticised in the fierce US debate mentioned above, as it excludes people with 'some difficulty'. The Washington Group team defended this on the grounds that in cognitive testing, 'a lot of difficulty' is more consistently interpreted across countries than 'some difficulty' (Mont *et al.*, 2024). However, this debate primarily applies to the Washington Group Short Set and the scoring for the longer question sets that cover a greater range of domains of impairments and activity limitations are more complex.

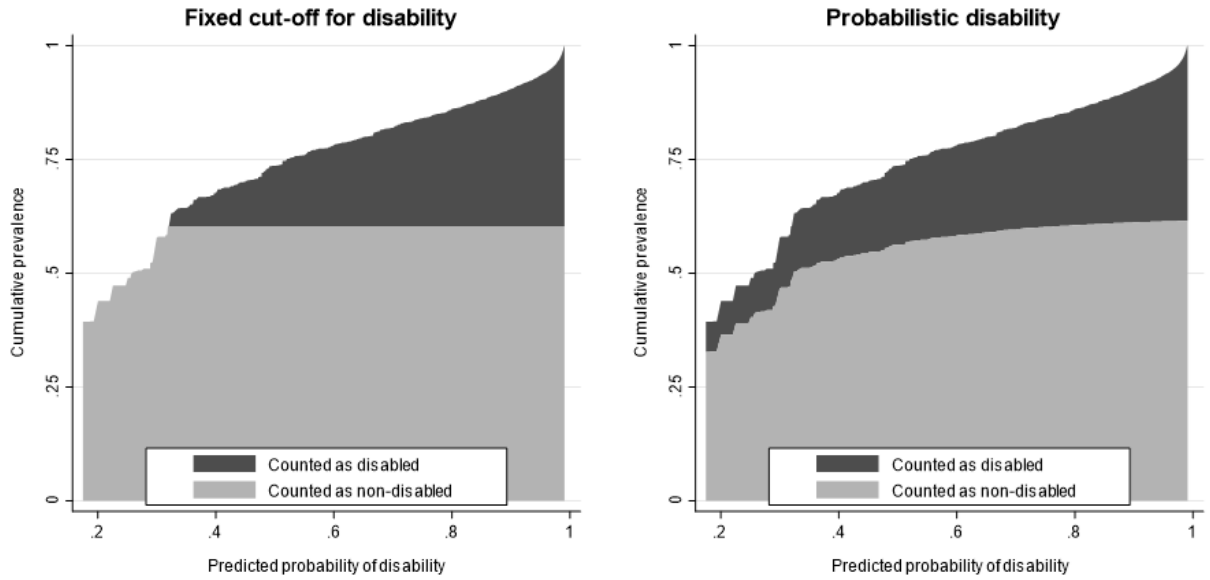
Whatever the exact approach, in all these cases, they use a fixed cut-off on the scale. However, there is an alternative – to *probabilistically* assign each person to disability, based on their estimated probability of reporting a disability. This means each respondent has a chance of being classified as having a disability, but the probability rises as respondents report more limitations. For example:

- Imagine two people: person A is estimated as having a 33% chance of reporting a disability, while person B is estimated as having a 67% chance of reporting a disability. (These percentages come from the predicted disability scale that we explain above).
- We then randomly allocate each person to being either 'disabled' or 'not disabled', based on these probabilities. (This is a bit like rolling the dice – person A's dice has 'disability' written on two of its six sides while person B's dice has 'disability' written on four sides). In the first replication, both happen to have a disability; in the second replication, neither of them; and in the third replication, person A has no disability, while person B has a disability.
- If we repeat this 100 times, then the randomness in these rolls of the dice is likely to average out: person A will probably have a disability in about 33 of these replications, while person B will have a disability in about 67 replications.

If we use this probabilistic disability measure for everyone in the sample, we can then use these 100 replications to estimate the disability employment gap. Roughly speaking, the mean disability employment gap is the average of the gap across these 100 replications; and the 95% confidence interval around the disability employment gap can be found from the average of the upper and lower bounds of the 95% confidence interval across these 100 replications. (There's also an alternative way of doing this that we explain in more detail in Appendix A5).

We can see the impact of this in Figure 3.2. For the fixed cut-off for disability (left-hand panel), people with the lowest predicted disability scores are never allocated to the 'disability' category, whereas people over the threshold are always treated as 'disabled'. In contrast, for probabilistic disability (right-hand panel), there is no threshold. Instead, people with the lowest level of disability still have some chance of being allocated to the 'disability' group, in line with their empirically observed probability of reporting a disability.

Figure 3.2. Visualising probabilistic disability vs. a fixed cut-off on a predicted disability scale



Source: Authors' analysis of SHARE-ELSA-HRS data – see text for explanation.

Why would we want a probabilistic disability measure?

While it has not been used before, there are two reasons why a probabilistic disability measure might be better than a fixed threshold: firstly, it is more comparable and secondly, it focuses more on people that policy makers are likely to care about.

Firstly, fixed-threshold disability measures are generally only as comparable as their *least* comparable component. In our case, most people reporting an impairment have a disability scale score that is above the threshold to classify them as 'disabled' (Table 3.1). So even if we have ten comparable impairment measures and only one less comparable one, the least comparable measure will generally be sufficient to drive differences in disability prevalence. In contrast, probabilistic disability measures are less affected by single measures – they use information on the entire disability scale, so the contribution of a single measure is restricted to the size of its weight (compared to the weights of all the other measures).

We can see this in Table 3.1. For Persons B, C and D, using a fixed disability threshold, reporting a limitation on an incomparable measure is sufficient to change them from being classified from the group 'non-disabled' to the group 'disabled'; once this has happened, then whether they report other limitations 1, 2, and 3 is irrelevant, and these situations are identical. Thinking probabilistically, though, the incomparable measure only changes the probability of reporting a disability by 5%, Person D is treated very differently to Person B based on their patterns of responding to other questions. In most cases, then, probabilistic measures will therefore be much more robust to comparability problems in, say, 1-2 indicators within the disability scale.

Secondly, fixed-threshold disability measures focus their attention primarily on people just above the threshold for disability – which is not helpful for policy. Imagine a threshold for disability of 32%, and three people with predicted probabilities of disability of 31%, 32%, and 99%. The first person (p=31%) is not classified as having a disability because they fall just underneath the fixed cut-off – any employment barriers that they face will serve to *reduce* the disability employment gap, because they bring down the employment rate among people categorised as 'non-disabled'. The second and third person are counted as 'disabled', and their employment status is given equal weight in calculating the disability employment

gap, even though one of them is much more likely to report a disability than the other. There are analogous problems in the Washington Group scoring approach, when, e.g., someone who reports a lot of anxiety most days is treated as 'not disabled' (because the threshold is set as feeling this 'every day'), but once someone has reported any limitation of a given severity, their other responses are ignored.

Table 3.1. The robustness of fixed vs. probabilistic disability measures to the incomparability of single items of the disability scale

	Disability weight[1]	Person A	Person B	Person C	Person D
Incomparable limitation	0.25	N	Y	Y	Y
Other limitation 1	0.25	N	N	Y	Y
Other limitation 2	0.25	N	N	N	Y
Other limitation 3	0.25	N	N	N	Y
Logit scale score[1]	n/a	-1	-0.75	-0.5	0
Probability of disability	n/a	27%	32%	38%	50%
Disability from fixed threshold (p ≥ 30%)	n/a	N	Y	Y	Y
Disability from Washington Group scoring[2]	n/a	N	Y	Y	Y

Note: The table shows a hypothetical example.
1. Disability weights refer to scores on a logistic scale (from a hypothetical logistic regression model where the intercept on the logit scale is -1.0). The probability of disability in this simplified case is therefore the inverse of the logit scale score.
2. 'Washington Group scoring' simply means that someone is treated as having a disability if they report any of these limitations.[33]

Probabilistic disability measures are therefore better in two respects – they pay more attention to people just below the fixed threshold for disability (rather than ignoring them), and they give much higher weights to people likely to have a disability. This may well change the comparative picture because as we noted above, countries can have low disability employment gaps for people just above the disability threshold, but high disability employment gaps for people with a high chance of reporting a disability (as mentioned before, the country-level correlation between these two disability employment gaps is only 0.48; see Appendix B). Probabilistic disability measures are therefore quite intuitive: they are a single measure that captures the degree of disability employment disadvantages, weighted to different groups according to how likely they are to report a disability.

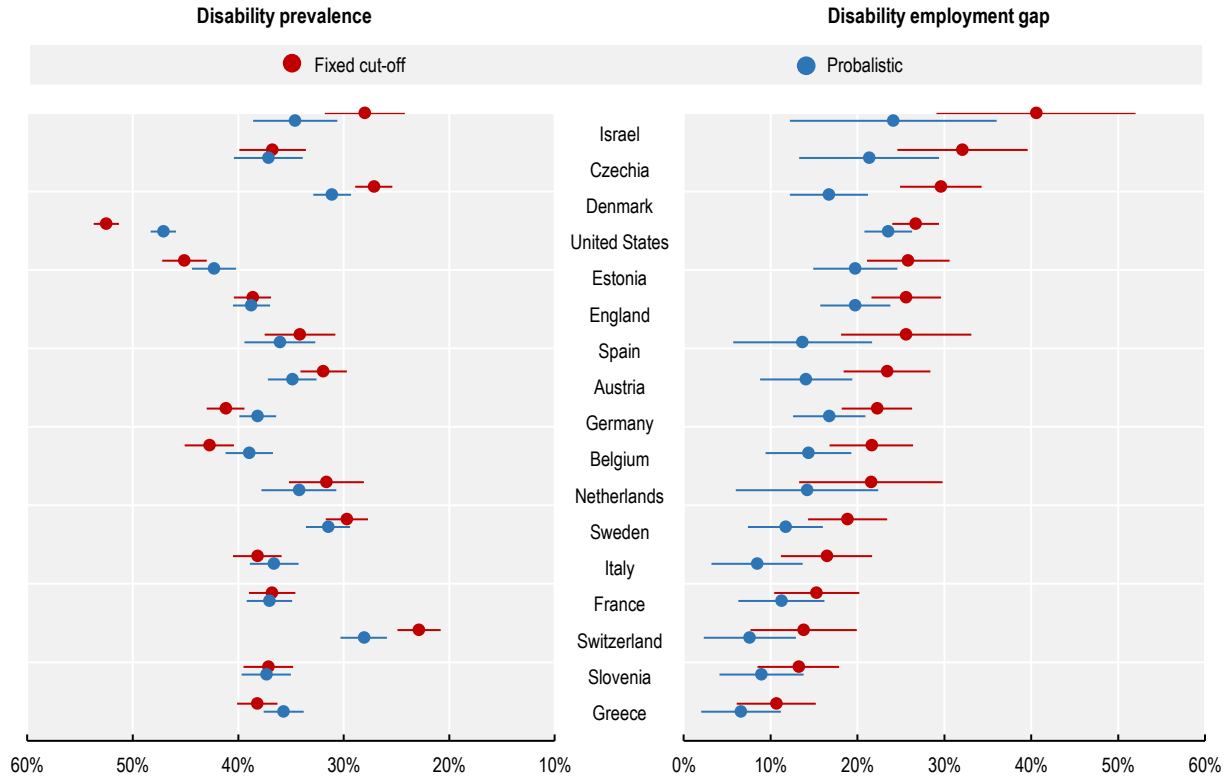
What difference do probabilistic measures make to our results?

The effect of using a probabilistic disability measure on our results is shown in Figure 3.3. Looking first at disability prevalence, in most cases the ranking is roughly similar between measures (left panel). This may be surprising, as the fixed threshold measure is very sensitive to the exact level of the threshold, and countries do differ in the number of people just above and just below the threshold.

However, it turns out that this has relatively little impact on the comparative picture because (given the prevalence of disability) the distribution of disability in each country is similar. That said, the probabilistic measure does tend to produce lower variability in disability between countries, as seen most markedly for the United States (which falls from 53% to 47%) and Switzerland (which rises from 23% to 28%). Using the probabilistic approach as opposed to using fixed cut-offs, the comparative differences in disability prevalence seem *prima facie* more plausible (e.g. comparing these to the morbidity estimates from the WHO's Global Burden of Disease Study).[34]

Despite this, the disability employment gap across countries varies much more (right panel), reflecting the fact that countries can differ markedly in their employment gaps for medium vs. high-level disability (as discussed above). In general, the disability employment gap is lower using probabilistic allocation vs. fixed cut-offs. However, the extent of this varies by country: in the United States there is little difference, whereas in Czechia and Denmark the disability employment gap is about 15 percentage points smaller when using probabilistic allocation. Overall, while the ranking changes in some respects, we still find that Switzerland, Greece, and Slovenia have the smallest gaps when using either approach.

Figure 3.3. The impact of using a probabilistic vs. a fixed threshold on disability prevalence and employment gaps among 50-69-year-olds, using a predicted disability scale



Source: Source: Authors' analysis of SHARE-ELSA-HRS data. Estimates based on average marginal effects after logistic regression models, setting age and sex to the all-country/all-survey means. Probabilistic results are based on 1,000 replications (in which each person is considered 'disabled' or not according to their probability of reporting a disability). Exact figures and confidence intervals are given in Appendix B.

Conclusion to Chapter 3

The best solution to the limited comparability of single-item disability measures is to use specific questions on impairments and activity limitations and combine these into a summary measure. We have explained several decisions within this process and reviewed their impact.

- Disability prevalence is affected by these methods – the wide (implausible) comparative differences in single-item activity-limiting disability narrow considerably when using a probabilistic measure, such that countries differ from each other less than half as much.[35] Disability prevalence in some countries is particularly affected, including the Netherlands and Greece.
- Disability employment gaps are consistently smaller when using the probabilistic approach than using single-item activity-limiting disability (a byproduct of the approach itself). While there are

some differences in the country ranking – Sweden and Switzerland do better, Spain and the Netherlands do worse – in general the comparative picture is roughly similar.

Note that there is no logical requirement for using these methods to have these effects; this is just one example of the difference they may make. If we consider a further example using the EHIS data (see Chapter 4), we find very large differences in disability prevalence *and* the disability employment gap between these methods. This confirms the conclusions to Chapter 3: there are times in which using these methods will produce similar results, and other times they will produce quite different results.

4 Which countries do best on disability employment gaps?

Previous international studies have interpreted country differences with caution because it was assumed that these differences may reflect methodological limitations rather than real patterns. Using more robust (albeit imperfect) methods to measure disability and employment reveal two cross-national patterns. First, countries with lower employment rates often have lower disability employment gaps. Second, some countries buck this trend, combining higher disability employment rates with low disability employment gaps. Exploring the reasons behind these patterns could drive future research, which may in turn help improve inclusion for people with disability into work.

Disability employment gaps have regularly been used in international academic research and policy reports – yet this has always been done hesitantly. OECD reports have often shown disability employment gaps across countries, but usually to frame discussions, rather than to drive conclusions. The authors of a major EU comparative study similarly note that untrustworthy disability employment gap comparisons often have an 'expressive function' – *"even incomplete data or data which do not lend themselves to comparative purposes can still serve a helpful public education and public policy function"* (Priestley and Grammenos, 2021). In other words, comparative disability employment gaps have been used to show that people with disability have lower chances of working in most countries, but researchers did not trust the international comparisons sufficiently to probe them further.

This has started to change, with greater numbers of academic papers examining which countries have the lowest disability employment gaps (Geiger *et al.*, 2019; Geiger *et al.*, 2017; Gugushvili *et al.*, 2023; Reinders Folmer *et al.*, 2020; van der Zwan and de Beer, 2021). There are also times that policymakers seem to have noticed particularly high/low-performing countries, with, e.g., the European Commission's country reports to Belgium, Bulgaria, Cyprus, Czechia and Germany noting their relative disability employment gaps, at least in passing (Priestley and Grammenos, 2021). Still, in general it seems that there is some hesitancy about the robustness of disability employment gap comparisons, with major reviews of effective policies around disability and work paying little attention to them.

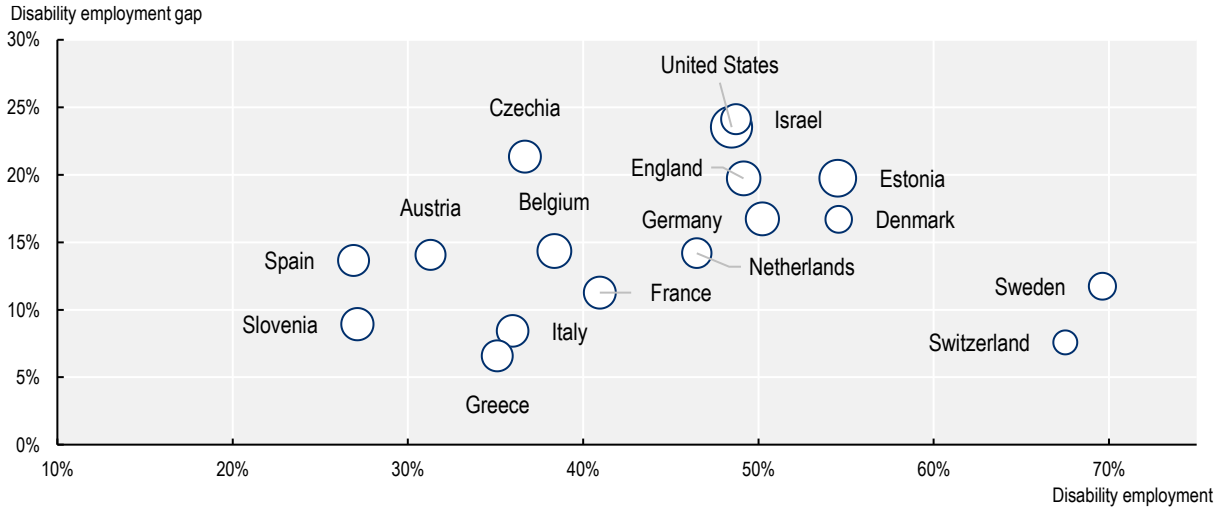
However, in this report, we have shown that more robust estimates of the disability employment gap are possible. We focus here on the results from the probabilistic disability measure, which we argue in Chapter 3 is likely to produce the most robust comparisons of disability employment across time and space.

Most robust results (using SHARE-ELSA-HRS)

In Chapter 3 we showed disability prevalence and disability employment gaps for people aged 50-69 but our focus was primarily on the difference between methodological approaches. In now focusing on the substance of this picture, it is important to further consider *absolute* levels of disability employment: low disability employment gaps may partly reflect low employment rates *per se*, which are lower for older ages in Southern Europe. Rather than arguing that either disability employment gaps or disability employment rates should be discarded at the expense of the other (Gugushvili *et al.*, 2023), we instead present these simultaneously in Figure 4.1.

The figure shows that countries that have higher disability employment gaps tend to have higher absolute employment rates for people with disability – that is, people with disability are doing better in countries like Denmark and Estonia, but people without disability are doing better still. Countries like Italy and Greece have low disability employment gaps, but this reflects their low employment rates overall – people with disability are still relatively unlikely to be in work. Against this general trend though, Switzerland and Sweden stand out: they have very high levels of employment among people with disability, and low disability employment gaps.

Figure 4.1. Disability employment rates vs. employment gaps for 50-69-year-olds (SHARE data)



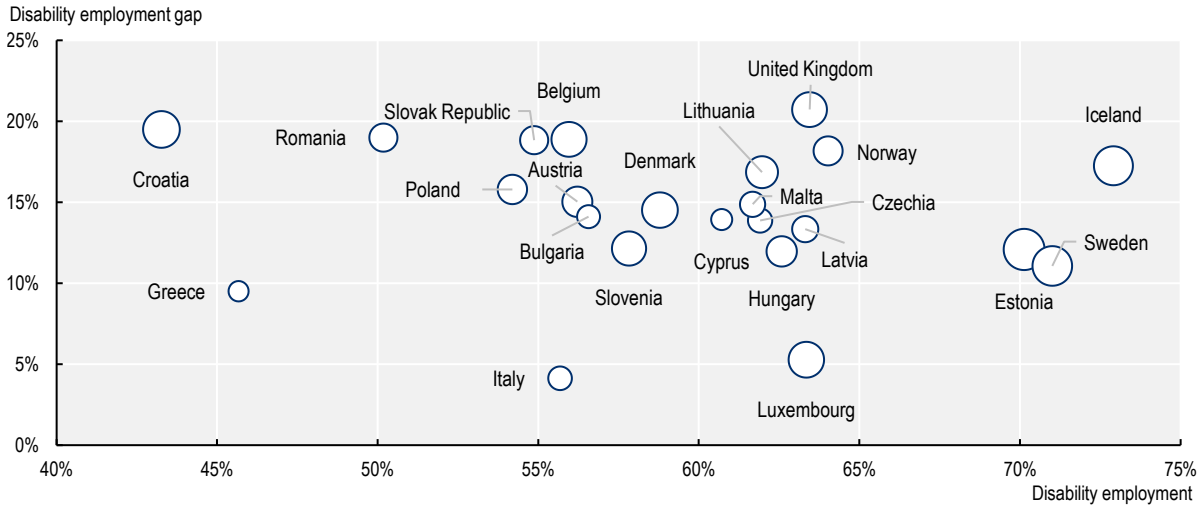
Source: Authors' analysis of SHARE-ELSA-HRS data. Bubble width represents the prevalence of disability (not area). Disability employment estimates based on average marginal effects after logistic regression models, setting age and sex to the all-country/all-survey means. Probabilistic results are based on 1 000 replications (in which each person is considered 'disabled' or not according to their probability of reporting a disability). Exact figures and confidence intervals are given in Appendix B

It is interesting to note that this picture is similar – but with some differences – to using less robust, single-item disability measures (see Appendix Figure B3). Some aspects are the same: Sweden and Switzerland are in the bottom-right corner (with the highest disability employment rates, and relatively low gaps); while Israel and Czechia are near the top-centre (with high gaps and middling employment rates). This also confirms patterns that may have been thought to reflect methodological limitations; for example, countries like Italy and Greece have low disability employment gaps, despite disability inclusion policies that lag behind other countries (although this comes alongside low disability employment *per se*). At the same time, some countries move significantly – Israel is no longer an outlier in the disability employment gap; and the link between disability employment rates and disability employment gaps is much more noticeable when looking at probabilistic disability.

All-age comparisons (using EHIS)

It is also possible to use the probabilistic approach to look at disability employment for the full working-age population (that is, people aged 20-69 rather than those aged 50-69) using the EHIS survey. A smaller set of measures are available in EHIS than for SHARE above; we here construct our predicted disability scale using only five sets of measures, covering hearing, seeing, difficulty walking, bodily pain, and mental health (see Appendix D2). We present this with caveats: as well as the limited number of limitations and health measures, we have already said that this survey is less comparable than SHARE (with different survey modes, different sampling approaches, and sometimes even different question wording, all of which will influence the results). Nevertheless, we present these tentative results from EHIS as it is important to consider the full age range of the working-age population, and this is shown in Figure 4.2.

Figure 4.2. Disability employment rates vs. employment gaps for all working-ages (EHIS data)



Source: Authors' analysis of EHIS data. Bubble width represents the prevalence of disability (not area). Estimates based on average marginal effects after logistic regression models, setting age and sex to the all-country/all-survey means. Probabilistic results are based on 1 000 replications (in which each person is considered 'disabled' or not according to their probability of reporting a disability). Exact figures and confidence intervals are given in Appendix B.

There are some differences in the patterns above – Italy and Greece perform similarly for older people in ELSA-SHARE but Italy performs much better than Greece for all ages in EHIS, while Estonia and Czechia are better-performing for all ages in EHIS compared to older people in SHARE. (Note that the figure does not display statistical uncertainty, so some of these may reflect random fluctuations in the data, rather than genuine differences by age or survey; confidence intervals are shown in Appendix B2). Still, overall, this shows many similarities to the picture above, with Sweden again an outlier in combining exceptionally high disability employment with a relatively low disability employment gap. The lowest disability employment gaps are found in Southern European countries of Greece and Italy, in both cases combined with a relatively low disability employment rate. Luxembourg is a noticeable case with a low gap despite an above-average rate.

While not our focus in this chapter, it is worth stressing that these results are noticeably different to the picture that we get from the conventional single-item activity-limiting disability measure in EHIS (Appendix Figure B4). Using the conventional measure, we find that disability prevalence varies fivefold (from 7% in Greece to 39% in Latvia), compared to only twofold when using the probabilistic measure – the range of the probabilistic measure seems more plausible *a priori*, and closer to the best available evidence from the Global Burden of Disease study (see footnote 34). The disability employment gap results also change noticeably, with e.g. the United Kingdom moving from being a medium-performing to low-performing country, and Sweden and Luxembourg becoming better-performing than Latvia. This again confirms that the newer methods that we present here can result in substantively different conclusions to conventional methods – we return to this issue in Chapter 5.

All-age trends (using EHIS)

It is also possible to look at trends over time by comparing data from EHIS wave 2 (2013-15) and EHIS wave 3 (2018-20). Here the main issue is the *consistency* of methods over time, but this is difficult to establish, and there are relatively few published studies that seek to compare EHIS over time (we can find no published Eurostat analyses of trends, with one exception (we can find no published Eurostat analyses

of trends, although see Arias-de la Torre *et al.*, 2023). The value of repeated coordinated national surveys is considerably enhanced by clearer information on comparability over time, as we return to in Chapter 5.

As an initial step in this direction, we exclude countries that seem to completely change survey mode between waves, and flag countries where there were some changes in survey mode (Croatia, Estonia, Hungary, Lithuania, Malta, Sweden, United Kingdom). The results comparing probabilistic vs. single-item activity-limiting disability measures are shown in Appendix Figure B5, which show:

1. Single-item activity-limiting disability in EHIS is not comparable over time. There are enormous, implausible drops in single-item disability between waves – in five countries it falls by 5-10 percentage points, and in another four it falls by more than 10 percentage points (Greece, Lithuania, Luxembourg, Slovak Republic). In comparison, when we use probabilistic disability there are relatively few changes, the changes that are seen are much smaller, and they are as likely to show rises in disability as falls.
2. In almost no country do we see a statistically significant change in the disability employment gap when using probabilistic measures. (Conversely, we see rises in employment gaps in many countries if we use the implausible single-item activity-limiting disability measures).

To reiterate: we know that there are issues in the comparability of EHIS, which means that these results must be read with caution. Still, this again shows the value of the probabilistic approach, which produces more plausible trends in disability prevalence and disability employment gaps.

Conclusions on which countries do best

Previous international studies have interpreted country differences with caution because it was assumed that these differences may reflect methodological limitations rather than real patterns. Using more robust (albeit imperfect) methods to measure disability and employment reveals two cross-national patterns:

1. First, when using these improved methods, countries with lower employment rates often have lower disability employment gaps.
2. Second, some countries buck this trend, combining higher disability employment rates with low disability employment gaps. Testing if this is borne out with other data sources, and exploring the reasons behind these patterns could drive future research, which may in turn help improve inclusion for people with disability into work.

Having explained the different ways of measuring the disability employment gap, and illustrated these with the best available data, we now need to draw this together. The final chapter therefore concludes with our recommendations on: what is the best way of estimating the disability employment gap?

5 Conclusions for comparing disability employment gaps in future across countries and over time

Unfortunately, more robust disability measures are constrained by the limited availability of detailed comparative data on impairments and activity limitations. This is not because validated question sets do not exist, but because these have not been used in an input-harmonized survey across OECD countries. Researchers, policy makers, and statistical agencies are therefore encouraged to work together to create better data for international comparisons of disability across the full working-age population.

Whatever techniques are used, researchers need to be aware of the challenges of comparability in measures of the disability employment gap. International comparisons and comparisons within countries over time are both difficult and crucial for policymaking; they need to be done with care.

In this paper, we have drawn attention to problems with the most common measure of the disability employment gap, which use single-item activity-limiting disability questions. These measures do have some value, as they are simple, and focus on people whose participation in society is affected by their impairments and activity limitations, which is the group that policymakers are usually concerned about. They are also short, and usually the only realistic option within multi-purpose surveys (e.g., labour force or household income surveys). However, they are not only unreliable but are affected by the very thing that we are evaluating (disability inclusion) – which means that while they are useful for some purposes, they are not robust for policy evaluations, nor for comparisons over time or across countries.

In its place, we have discussed different ways of getting more robust comparisons over time and across countries. In this final chapter, we draw together our recommendations for both analysts and data producers, and the policymakers that both ultimately need to satisfy. It is clear from recent US controversies over disability measurement that headline statistical measures need to be supported by key stakeholders (ideally including researchers, international organisations, and organisations of people with disabilities). By 'recommendations' in this concluding section, we therefore do not mean that the following proposals should be unilaterally imposed; instead, we are putting forward proposals that we believe would provide better evidence, and therefore merit consideration by others within this wider debate.

Recommendations for analysts

Using single-item activity-limiting disability measures

Even when using conventional single-item activity-limiting disability measures, there are ways that we can make our comparisons more robust (see Chapter 2):

- One option is to use **more-comparable studies**. Coordinated national studies are invaluable for showing that people with disability face employment barriers in all countries, and providing large, high-quality datasets for research. But for robust comparisons of countries, the flexibility that is necessary for coordinated national studies means that we cannot be sure if the differences that we see reflect genuine patterns or the results of varying methodological choices. International studies that harmonize the methods of data collection are therefore the best source for robust comparisons.
- Another option is to use a further measure, the **'prevalence-adjusted disability employment gap'**. This is simply the conventional disability employment gap multiplied by the prevalence of disability (and shows the share of the population who are potentially prevented from working due to disability). However, it can still make our comparisons over time and place more robust. Sometimes this is because the prevalence-adjusted gap is more likely to be unbiased. But more commonly, similar patterns in the conventional gap and the prevalence-adjusted gap give us more confidence in our results – and where they show different patterns, this shows that one or both measures may be untrustworthy.

The takeaway message is that the conventional gap and the prevalence-adjusted gap are useful complements to each other – they highlight different things that policymakers need to know; and where they show different patterns, they alert us to the limitations of these single-item measures of disability.

Using multi-item disability scales

However, the most robust comparisons require better measures of disability. It is widely recognised that the best way of doing this is to use a series of specific questions on impairments and activity limitations, which are answered reasonably similarly wherever and whenever they are asked, which are then combined into a single scale. While there is some disagreement over exactly how to construct such a scale, we are convinced that the underlying questions should be focused on impairments/activity limitations, as these

are more likely to be interpreted consistently across different times and places (see Chapter 3). This means that we are measuring disability in terms of people who are *potentially disadvantaged* in the environments that we see across OECD countries (or in the ICF's terms, as 'capacity' rather than 'performance'). The analysis is still based on the ICF/social model of disability; but to uncover disabling social environments, we fundamentally need a measure of disability that is measured consistently across different environments.

While there are many examples of creating disability scales, researchers often seem unaware of the limitations of their methods, and the different choices available that might produce more robust results. In Chapter 3 we recommend that:

- **The items included in the scale should be comparable and comprehensive:** they should focus on measures that are (i) likely to be comparable across time and place; and (ii) cover all the dimensions of impairments and activity limitations likely to be work-disabling in OECD countries. These can sometimes be in tension, as some, e.g., mental health or pain, are inherently difficult to compare across countries – but they must be included somehow, as their omission means that we fail to understand the realities of work disability.
- **The weights for each item in the scale should be produced using the 'predicted disability' approach** – that is, by weighting each item according to how strongly it predicts single-item activity-limiting disability. On paper, this is a much more sensible approach than weighting each item according to how strongly it is associated with the other items in the scale (as that would mean that, e.g., mental health is given a low weight if most measures are about physical impairments). In practice, though, different weights will only occasionally affect comparisons.
- **The scale should be turned into a binary disability category using a probabilistic approach**, rather than a fixed threshold. That is, we estimate the predicted probability of each person reporting a disability given the impairments and activity limitations they report and run a series of replications where each person is treated as '(not) having a disability' in each replication based on this probability. This is likely to be more comparable across time and place because it is less sensitive to any individual question. It also pays more attention to people just below an arbitrary fixed threshold for disability (who are otherwise ignored) and gives higher weights to people likely to have a disability.

Alternatively, if researchers prefer to keep a fixed threshold, then it is important to look at different levels of this threshold and to look at employment gaps for people with moderate vs. high chances of reporting a disability, which may be quite different.

Of these three stages, it is the third (the probabilistic approach) that makes the most difference to the cross-national comparison in our example analysis. However, across all of our analyses in Chapters 2-4, we find that each of these methodological choices sometimes result in similar findings and sometimes result in different ones – there is no way of knowing *a priori* which methodological choices will matter.

Overall, our view is that the most robust basis for comparing disability employment is to use a predicted disability scale, based on comparable and comprehensive measures, which is turned into a binary 'disability' measure via the probabilistic approach. But what is most important is that we are aware of the impact of the choices we make, and that whatever decisions we take, we move beyond using single-item activity-limiting disability measures. However, because better data are needed for this, it is also necessary to make recommendations to data producers.

Recommendations for data producers

Unfortunately, the possibility of using these proposed, more robust disability measures is constrained by the limited availability of detailed comparative data on impairments and activity limitations. Like other researchers, we primarily use the Global Aging Surveys, which only cover people aged 50 and over. (We have shown that it is also possible to use this approach on EHIS data, but the number of measures is

limited, the documentation of within-country methodological changes could be improved, and there are several methodological differences between countries that may influence the results.)

Comprehensive, validated question sets on impairments and activity limitations do exist (see Box 3.1), but to date these have not been used in an input-harmonized survey across OECD countries. However, given the survey space required, it is simply not plausible that these batteries of questions could be included in all surveys that currently contain single-item measures such as GALI. The most cost-efficient way forward is therefore to supplement existing surveys with periodic '*calibration surveys*' that provide more robust estimates of disability and employment.[36] These surveys should be purpose-designed to robustly monitor disability inclusion: they should be consistent across time/place and include sets of questions on impairments and activity limitations that balance both coverage and comparability (see Chapter 3).

In the short-term, individual national governments may test the value of such calibration surveys for understanding trends within their country. The aim is not to replace the short, useful disability screening instruments used in various surveys (which in countries like Canada and the United Kingdom has been standardised recently across all population surveys), but rather to invest in new data collection for the specific purpose of making comparisons over time, which thereby provides a check on the interpretation of the more-regular, but less-robust, measures in other surveys.

Looking comparatively, there are two ways an international calibration survey could come about. One approach would be to create a new international survey of working-age disability, which is extensively resourced to enable a sufficient sample size of people with disability, while maintaining input-harmonization on the level of the ESS or SHARE. The viability of this approach obviously depends on funding. The other approach would be to build such measures into existing efforts of coordinated national surveys, e.g. for EHIS wave 5 in 2031 – much as such processes must balance numerous conflicting demands, and this will only occur many years into the future. If this is the case, then as well as making efforts to maximise methodological harmonization within a framework of national flexibility, the transparency of divergent methodological choices is also crucial.

Either way, we strongly encourage researchers, policy makers, and statistical agencies to work together to create calibration surveys, to provide better international comparisons and trend analyses of disability across the full working-age population.

Conclusion

Whatever data and analytical techniques are used, researchers and policy makers need to be aware of the challenges of comparability in measures of the disability employment gap. These challenges are greatest when using single-item activity-limiting disability measures – but even when using multi-item disability scales, there may be measures within the scales that are interpreted differently in different times and places. Beyond this, there are also differences in response rates and response biases, challenges in, e.g., accounting for people that are excluded from most household surveys (Binswanger *et al.*, 2009; Kaspar *et al.*, 2023), and differences in whether people who struggle to complete the survey can be counted through proxy responses (Weir *et al.*, 2011). International comparisons and comparisons over time within countries are both difficult and crucial for policymaking; they need to be done with care.

References

Altman, B.M. (2016), International Measurement of Disability: Purpose, Method and Application, *Social Indicators Research Series 61*: Springer.

Arias-de la Torre, J., et al. (2023), 'Prevalence and variability of depressive symptoms in Europe: update using representative data from the second and third waves of the European Health Interview Survey (EHIS-2 and EHIS-3)', *The Lancet Public Health*, 8: 11, e889-e898.

Arora, V.S., et al. (2015), 'Data Resource Profile: The European Union Statistics on Income and Living Conditions (EU-SILC)', *International Journal of Epidemiology*, 44: 2, 451-461.

Banks, J. and Smith, J.P. (2012), 'International Comparisons in Health Economics: Evidence from Aging Studies', *Annual Review of Economics*, 4, 57-81.

Baumberg, B., Jones, M. and Wass, V. (2015), 'Disability prevalence and disability-related employment gaps in the UK 1998–2012: Different trends in different surveys?', *Social Science & Medicine*, 141, 72-81.

Berger, N., et al. (2015), 'Assessing the validity of the Global Activity Limitation Indicator in fourteen European countries', *BMC Medical Research Methodology*, 15: 1, 1.

Berthoud, R. (2011), Trends in the Employment of Disabled People in Britain, *ISER Working Paper No 2011-03*: Institute of Social and Economic Research (ISER), University of Essex. http://www.iser.essex.ac.uk/publications/working-papers/iser/2011-03

Bickenbach, J., et al. (2023), 'The human functioning revolution: implications for health systems and sciences', *Frontiers in Science*, Volume 1 - 2023.

Bickenbach, J.E. (2011), 'Monitoring the United Nation's Convention on the Rights of Persons with Disabilities: data and the International Classification of Functioning, Disability and Health', *BMC Public Health*, 11: 4, S8.

Bickenbach, J.E. (2019), 'The ICF and its relationship to disability studies', in N. Watson, A. Roulstone and C. Thomas (eds.), *Routledge handbook of disability studies [2e]*: Routledge.

Biermann, J. and Pfahl, L. (2021), 'A Global Monitoring Practice in the Making : Disability Measurement for UN Sustainable Development Goal 4 on Inclusive Education', *Österreichische Zeitschrift für Geschichtswissenschaften*, 31: 3, 192-213.

Binswanger, I.A., Krueger, P.M. and Steiner, J.F. (2009), 'Prevalence of chronic medical conditions among jail and prison inmates in the USA compared with the general population', *Journal of Epidemiology and Community Health*, 63: 11, 912-919.

Böheim, R. and Leoni, T. (2015), Disability Policies across Europe: Reforms and Employment Outcomes for Workers Aged 50+, *NBER Disability Research Center Paper No. NB 15-08B*, Cambridge, MA: National Bureau of Economic Research (NBER). http://www.nber.org/aging/drc/papers/odrc15-08b

Börsch-Supan, A. (2017), Survey of Health, Ageing and Retirement in Europe (SHARE) Waves 1-6. Release version: 6.0.0.: SHARE-ERIC. http://dx.doi.org/10.6103/SHARE.w6.600

Bound, J. (1991), 'Self-reported versus objective measures of health in retirement models', *Journal of Human Resources*, 26: 1, 106-138.

Bowling, A. (2005), 'Mode of questionnaire administration can have serious effects on data quality', *Journal of Public Health*, 27: 3, 281-291.

Burkhauser, R.V., Houtenville, A.J. and Tennant, J.R. (2014), 'Capturing the Elusive Working-Age Population With Disabilities: Reconciling Conflicting Social Success Estimates From the Current Population Survey and American Community Survey', *Journal of Disability Policy Studies*, 24: 4, 195-205.

Casebolt, M.T. (2021), 'Availability and quality of global disability data: A commentary on the Demographic and Health Surveys', *Disability and Health Journal*, 14: 1, 100972.

Cases, C. (2021), 'IESS: Europe Harmonises its Social Statistics to Better Inform Public Policies', *Courrier des*

statistiques, 3.
Castro-Costa, E., et al. (2008), 'Ascertaining late-life depressive symptoms in Europe: an evaluation of the survey version of the EURO-D scale in 10 nations. The SHARE project', *International Journal of Methods in Psychiatric Research*, 17: 1, 12-29.
Cernat, A., Couper, M.P. and Ofstedal, M.B. (2016), 'Estimation of Mode Effects in the Health and Retirement Study Using Measurement Models', *Journal of Survey Statistics and Methodology*, 4: 4, 501-524.
Choi, S.W., Gibbons, L.E. and Crane, P.K. (2011), 'lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations', *Journal of Statistical Software*, 39: 8, 1 - 30.
Cieza, A., et al. (2015), 'The English are healthier than the Americans: really?', *International Journal of Epidemiology*, 229-238.
Cohen, F., et al. (2007), 'Immune Function Declines With Unemployment and Recovers After Stressor Termination', *Psychosomatic Medicine*, 69, 225-234.
Courtin, E., et al. (2015), 'Are different measures of depressive symptoms in old age comparable? An analysis of the CES-D and Euro-D scales in 13 countries', *International Journal of Methods in Psychiatric Research*, 24: 4, 287-304.
Crane, P.K., et al. (2006), 'Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques: DIFdetect and difwithpar', *Medical Care*, 44: 11, S115-S123.
Croezen, S., Burdorf, A. and van Lenthe, F.J. (2013), Agreement and disagreement in prevalence estimates of health between SHARE and other European population studies, *SHARE Working Paper Series 14-2013*. https://share-eric.eu/fileadmin/user_upload/SHARE_Working_Paper/WP_Series_14_2013.pdf
Croezen, S., Burdorf, A. and van Lenthe, F.J. (2016), 'Self-perceived health in older Europeans: Does the choice of survey matter?', *The European Journal of Public Health*, 26: 4, 686-692.
DWP and DH (2016), Work, Health and Disability Green Paper Data Pack, London: Department of Work and Pensions (DWP) and Department of Health (DH). https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/586723/work-health-and-disability-green-paper-data-pack.pdf [accessed 13/7/2017]
Elkasabi, M. (2020), 'Differences in Proxy-Reported and Self-Reported Disability in the Demographic and Health Surveys', *Journal of Survey Statistics and Methodology*, 9: 2, 335-351.
Eurostat (2010), 2010 EU Comparative Final Quality Report, Brussels: EC http://ec.europa.eu/eurostat/documents/1012329/6064601/2010+EU-FQR.pdf/3ccaea2b-7c1f-4b1c-a8e7-7af1a4284307 [accessed 13.05.2015]

Federici, S., et al. (2017), 'World Health Organization disability assessment schedule 2.0: An international systematic review', *Disability and Rehabilitation*, 39: 23, 2347-2380.
Fehr, A., et al. (2017), 'Health monitoring and health indicators in Europe', *Journal of Health Monitoring*, 2: 1, 3.
Fischer, F., et al. (2018), 'Measurement invariance and general population reference values of the PROMIS Profile 29 in the UK, France, and Germany', *Quality of Life Research*, 27: 4, 999-1014.
Geiger, B.B., van der Wel, K.A. and Tøge, A.G. (2017), 'Success and failure in narrowing the disability employment gap: comparing levels and trends across Europe 2002–2014', *BMC Public Health*, 17: 1, 928.
Geiger, B.B., et al. (2018), 'Assessing work disability for social security: international models for the direct assessment of work capacity', *Disability & Rehabilitation*, 40: 4, 2962-2970.
Geiger, B.B., Böheim, R. and Leoni, T. (2019), 'The growing American health penalty: International trends in the employment of older workers with poor health', *Social Science Research*, 82, 18-32.
Geiger, B.B. (2020), 'Has working-age morbidity been declining? Changes over time in survey measures of general health, chronic diseases, symptoms and biomarkers in England 1994–2014', *BMJ Open*, 10: 3, e032378.
Goddard, K.S. and Hall, J.P. (2025), 'Limitations of the Washington Group Short Set in capturing moderate and severe mobility disabilities', *Health Affairs Scholar*, 3: 2.
Gómez-Benito, J., et al. (2018), 'Differential Item Functioning: Beyond validity evidence based on internal structure', *Psicothema*, 30: 1, 104-117.
Grammenos, S. (2014), European comparative data on Europe 2020 & People with disabilities: Task 6: Comparative data and indicators, Brussels: Centre for European Social and Economic Policy on behalf of the Academic

Network of European Disability Experts (ANED).
Grammenos, S. (2025), European comparative data on persons with disabilities Equal opportunities, fair working conditions, social protection and inclusion: Analysis and trends - Data 2022: European Commission. https://op.europa.eu/o/opportal-service/download-handler?identifier=a2fce512-cf0a-11ef-be2a-01aa75ed71a1&format=pdf&language=en&productionSystem=cellar&part=
Groce, N.E. and Mont, D. (2017), 'Counting disability: emerging consensus on the Washington Group questionnaire', *The Lancet Global Health*, 5: 7, e649-e650.
Groves, R.M. and Peytcheva, E. (2008), 'The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis', *Public Opinion Quarterly*, 72: 2, 167-189.
Gugushvili, A., et al. (2023), 'No evidence that social-democratic welfare states equalize valued outcomes for individuals with disabilities', *Social Science & Medicine*, 339, 116361.
Hall, J.P., et al. (2022), 'Comparing Measures Of Functional Difficulty With Self-Identified Disability: Implications For Health Policy', *Health Affairs*, 41: 10, 1433-1441.
Hood, K., et al. (2012), 'Mode of data elicitation, acquisition and response to surveys: a systematic review', *Health Technology Assessment*, 16, 27.
Jacobs, P. (2024), 'U.S. Census Bureau scraps proposed changes to disability questions', *Science*, 7/2/2024. doi.org/10.1126/science.z9t17ru
Jadhav, A. and Weir, D. (2017), 'Widowhood and Depression in a Cross-National Perspective: Evidence from the United States, Europe, Korea, and China', *Journals of Gerontology Series B - Psychological Sciences and Social Sciences*.
Jones, M. and Wass, V. (2013), 'Understanding changing disability-related employment gaps in Britain 1998–2011', *Work, Employment & Society*, 27: 6, 982-1003.
Jones, M.K. (2006), 'Is there employment discrimination against the disabled?', *Economics Letters*, 92: 1, 32-37.
Jowell, R. (1998), 'How Comparative is Comparative Research', *American Behavioral Scientist*, 42: 2, 168-177.
Jürges, H. (2007), 'True health vs response styles: exploring cross-country differences in self-reported health', *Health Economics*, 16: 2, 163-178.
Kaspar, R., et al. (2023), 'Challenges and Benefits of Including the Institutionalized, Cognitively Impaired and Unable to Respond in a Representative Survey of the Very Old', *Survey Research Methods*, 17: 2, 111-129.
Kayess, R. and French, P. (2008), 'Out of Darkness into Light? Introducing the Convention on the Rights of Persons with Disabilities', *Human Rights Law Review*, 8: 1, 1-34.
Korkeila, K., et al. (2001), 'Non-response and related factors in a nation-wide health survey', *European Journal of Epidemiology*, 17: 11, 991-999.
Landes, S.D., Swenor, B.K. and Hall, J.P. (2024), 'Performance of the Washington Group questions in measuring blindness and deafness', *Health Affairs Scholar*, 2: 11.
Landes, S.D., et al. (2025), 'Comparative performance of disability measures', *PLOS ONE*, 20: 1, e0318745.
Lawson, A. and Beckett, A.E. (2021), 'The social and human rights models of disability: towards a complementarity thesis', *The International Journal of Human Rights*, 25: 2, 348-379.
Lee, L., et al. (2022), 'WHO Functioning and Disability Disaggregation (FDD11) tool: a reliable approach for disaggregating data by disability', *Archives of Public Health*, 80: 1, 249.
Lee, N., Cadogan, J.W. and Chamberlain, L. (2013), 'The MIMIC model and formative variables: problems and solutions', *AMS Review*, 3: 1, 3-17.
Lee, S., Mathiowetz, N.A. and Tourangeau, R. (2004), 'Perceptions of disability: the effect of self-and proxy response', *Journal of Official Statistics*, 20: 4, 671.
Loeb, M. (2016), 'International census/survey data and the short set of disability questions developed by the Washington group on disability statistics', in B. M. Altman (ed.), *International Measurement of Disability: Purpose, Method and Application*: Springer.
Madans, J., Mont, D. and Goodman, N. (2025), 'Letter to the Editor in response to Performance of the Washington Group questions in measuring blindness and deafness by Landes et al', *Health Affairs Scholar*, 3: 1.
Madans, J.H., Loeb, M.E. and Altman, B.M. (2011), 'Measuring disability and monitoring the UN Convention on the Rights of Persons with Disabilities: the work of the Washington Group on Disability Statistics', *BMC Public Health*, 11: 4, S4.
Maskileyson, D., Seddig, D. and Davidov, E. (2021), 'The EURO-D Measure of Depressive Symptoms in the Aging

Population: Comparability Across European Countries and Israel', *Frontiers in Political Science*, 3.
Miller, K., et al. (2011), 'Results of a cross-national structured cognitive interviewing protocol to test measures of disability', *Quality & Quantity*, 45: 4, 801-815.
Molden, T.H. and Tøssebro, J. (2010), 'Measuring disability in survey research: Comparing current measurements within one data set', *ALTER-European Journal of Disability Research/Revue Européenne de Recherche sur le Handicap*, 4: 3, 174-189.
Mont, D. (2007), Measuring disability prevalence, *Special Protection Discussion Paper No. 0706*, Washington, DC: World Bank.
Mont, D. (2019), 'How Are The Washington Group Questions Consistent With The Social Model Of Disability?'. https://www.washingtongroup-disability.com/wg-blog/how-are-the-washington-group-questions-consistent-with-the-social-model-of-disability-65/
Mont, D., et al. (2024), 'When Designing Disability Survey Questions, Align Measurement To Purpose: A Response To Landes et al.', *Health Affairs Forefront*, October 3, 2024. doi.org/10.1377/forefront.20240930.130625
Murray, C. and Chen, L. (1992), 'Understanding morbidity change', *Population and Development Review*, 18: 3, 481-503.
OECD (2003), Transforming Disability into Ability: Policies to Promote Work and Income Security for Disabled People, Paris: OECD. http://www.sourceoecd.org/
OECD (2010a), Sickness, Disability and Work: breaking the barriers. A synthesis of findings across OECD countries, Paris: OECD. http://www.oecd-ilibrary.org/social-issues-migration-health/sickness-disability-and-work-breaking-the-barriers_9789264088856-en
OECD (2010b), Sickness, disability and work: Improving social and labour-market integration of people with disability, Paris: Organisation for Economic Co-operation and Development (OECD). http://www.oecd.org/els/soc/46488022.pdf
OECD (2012), Sick on the Job? Myths and Realities about Mental Health and Work, Paris: Organisation for Economic Cooperation and Development (OECD). http://dx.doi.org/10.1787/9789264124523-en
OECD (2015), Fit Mind, Fit Job: From evidence to practice in mental health and work, Paris: OECD. http://dx.doi.org/10.1787/9789264228283-en
OECD (2022), Disability, Work and Inclusion: Mainstreaming in All Policies and Practices, Paris: OECD Publishing. https://doi.org/10.1787/1eaa5e9c-en
OECD (2023), Disability, Work and Inclusion in Italy: Better Assessment for Better Support, Paris: OECD Publishing. https://doi.org/10.1787/dc86aff8-en
Palmer, M. and Harley, D. (2011), 'Models and measurement in disability: an international review', *Health Policy and Planning*, 27: 5, 357-364.
Plessen, C.Y., et al. (2024), 'How Are Age, Gender, and Country Differences Associated With PROMIS Physical Function, Upper Extremity, and Pain Interference Scores?', *Clinical Orthopaedics and Related Research®*, 482: 2, 244-256.
Poterba, J., Venti, S. and Wise, D.A. (2013), 'Health, education, and the postretirement evolution of household assets', *Journal of Human Capital*, 7: 4, 297-339.
Priestley, M. and Grammenos, S. (2021), 'How useful are equality indicators? The expressive function of 'stat imperfecta' in disability rights advocacy', *Evidence & Policy*, 17: 2, 209-226.
Reinders Folmer, C.P., Mascini, P. and Van der Veen, R.J. (2020), 'Evaluating social investment in disability policy', *Social Policy & Administration*, 54: 5, 792-812.
Riumallo-Herl, C., et al. (2014), 'Job loss, wealth and depression during the Great Recession in the USA and Europe', *International Journal of Epidemiology*, 43: 5, 1508-1517.
Robine, J.-M. and Jagger, C. (2003), 'Creating a coherent set of indicators to monitor health across Europe: the Euro-REVES 2 project', *European Journal of Public Health*, 13: suppl_3, 6-14.
Rubio-Valverde, J.R., Nusselder, W.J. and Mackenbach, J.P. (2019), 'Educational inequalities in Global Activity Limitation Indicator disability in 28 European Countries: Does the choice of survey matter?', *International Journal of Public Health*, 64: 3, 461-474.
Sabariego, C., et al. (2015), 'Measuring disability: Comparing the impact of two data collection approaches on disability rates', *International Journal of Environmental Research and Public Health*, 12: 9, 10329-10351.
Sabariego, C., et al. (2016), 'Response to Madans et al. Comments on Sabariego et al. Measuring Disability:

Comparing the Impact of Two Data Collection Approaches on Disability Rates. Int. J. Environ. Res. Public Health, 2015, 12, 10329–10351', *International Journal of Environmental Research and Public Health*, 13: 1, 66.

Sabariego, C., et al. (2021), 'Measuring functioning and disability using household surveys: metric properties of the brief version of the WHO and World Bank model disability survey', *Archives of Public Health*, 79: 1, 128.

Sabariego, C., et al. (2022), 'Generating comprehensive functioning and disability data worldwide: development process, data analyses strategy and reliability of the WHO and World Bank Model Disability Survey', *Archives of Public Health*, 80: 1, 6.

Sabariego, C., et al. (2025), 'Can we still ensure no one is left behind by 2030? Demonstrating the potential of the implementation of the WHO Functioning and Disability Disaggregation Tool (FDD11) in existing survey platforms for disaggregating SDG indicators by disability', *Disability and Rehabilitation*, 47: 5, 1253-1265.

Saltychev, M., et al. (2021), 'Psychometric properties of 12-item self-administered World Health Organization disability assessment schedule 2.0 (WHODAS 2.0) among general population and people with non-acute physical causes of disability – systematic review', *Disability and Rehabilitation*, 43: 6, 789-794.

Shakespeare, T. (2013), 'Day of Recokining', *Discover Society*. https://archive.discoversociety.org/2013/11/05/day-of-reckoning/

Sousa, R.M., et al. (2010), 'Measuring disability across cultures — the psychometric properties of the WHODAS II in older people from seven low- and middle-income countries. The 10/66 Dementia Research Group population-based survey', *International Journal of Methods in Psychiatric Research*, 19: 1, 1-17.

Terwee, C.B., et al. (2021), 'International application of PROMIS computerized adaptive tests: US versus country-specific item parameters can be consequential for individual patient scores', *Journal of Clinical Epidemiology*, 134, 1-13.

Terwee, C.B. and Roorda, L.D. (2023), 'Country-specific reference values for PROMIS(®) pain, physical function and participation measures compared to US reference values', *Ann Med*, 55: 1, 1-11.

Todorov, A. and Kirchner, C. (2000), 'Bias in proxies' reports of disability: data from the National Health Interview Survey on disability', *Am J Public Health*, 90: 8, 1248-1253.

Üstün, T., et al. (2010), Measuring Health and Disability: Manual for WHO Disability Assessment Schedule (WHODAS 2.0), Geneva: World Health Organization. https://iris.who.int/bitstream/handle/10665/43974/9789241547598_eng.pdf?sequence=1

van der Wel, K.A., Dahl, E. and Thielen, K. (2012), 'Social Inequalities in "Sickness": Does Welfare State Regime Type Make a Difference? A Multilevel Analysis of Men and Women in 26 European Countries', *International Journal of Health Services*, 42: 2, 235-255.

van der Zwan, R. and de Beer, P. (2021), 'The disability employment gap in European countries: What is the role of labour market policy?', *Journal of European Social Policy*, 31: 4, 473-486.

Weir, D., Faul, J. and Langa, K. (2011), 'Proxy interviews and bias in cognition measures due to non-response in longitudinal studies: a comparison of HRS and ELSA', *Longitudinal and Life Course Studies*, 2: 2, 170-184.

WHO (2002), Towards a Common Language for Functioning, Disability and Health: The International Classification of Functioning, Disability and Health (ICF), *WHO/EIP/GPE/CAS/01.3*, Geneva: World Health Organization (WHO). https://cdn.who.int/media/docs/default-source/classification/icf/icfbeginnersguide.pdf

WHO (2011), *World report on disability*, Geneva: World Health Organization (WHO).

WHO and World Bank (2011), World Report on Disability, Geneva: World Health Organization (WHO).

WHO (2013), How to use the ICF: A Practical Manual for using the International Classification of Functioning, Disability and Health (ICF) [Exposure draft for comment], Geneva: World Health Organization (WHO). https://www.who.int/docs/default-source/classification/icf/drafticfpracticalmanual2.pdf?sfvrsn=8a214b01_4

WHO (2016), Global Health Estimates 2015: Disease burden by Cause, Age, Sex, by Country and by Region, 2000-2015, Geneva: World Health Organization (WHO).

Wirth, H. and Pforr, K. (2022), 'The European Union Statistics on Income and Living Conditions after 15 Years', *European Sociological Review*, 38: 5, 832-848.

Wise, D.A. (2017), Social Security Programs and Retirement Around the World: The Capacity to Work at Older Ages, Chicago: University of Chicago Press for the National Bureau of Economic Research (NBER). http://papers.nber.org/books/wise-22

Yin, N. and Heiland, F. (2015), Work Limitation Reporting and Disability Programs in Europe and the U.S., *Annual*

Meeting of the Population Association of America, April 30-May 2, San Diego. http://paa2015.princeton.edu/uploads/152896

Zumbo, B.D. (2007), 'Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going', *Language Assessment Quarterly*, 4: 2, 223-233.

Notes

[1] Different countries use different language – in the United Kingdom, people usually talk about 'disabled people', whereas elsewhere people usually talk about 'people with disability'. Because of the international focus of this paper, and in line with OECD practice, 'people with disability' is used here.

[2] The distinction between problems in body functions and in social roles is made by all key stakeholders, but different organisations use slightly different terms. For example, the UN Washington Group on Disability Statistics and Eurostat (see https://tinyurl.com/eurostatdisability) refer to problems in activity limitations as 'functional limitations'.

[3] In 2018 about 11% of the working-age population in Norway claimed a disability benefit (excluding the so-called Work Assessment Allowance; OECD, 2022). This is almost twice as high as the OECD average of around 6%.

[4] In the former case, the extra Norwegians reporting a disability are out of work, thus decreasing the employment rate of people with disability. In the latter case, the Norwegians that do *not* report a disability are likely to be those with less severe limitations who are more likely to work (for whom the decision to report a disability is more marginal), leaving the disability category to be more concentrated on those with lower employment chances.

[5] The EU-SILC has been used to estimate the disability employment gap by many researchers, including the OECD itself (2010b, 2022), but also the EU's Academic Network of European Disability Experts (Grammenos, 2014) and its follow-up European Disability Expertise network (Grammenos, 2025), and various academic researchers (van der Wel *et al.*, 2012), not least because it is one of the few surveys also including information on personal and household income as well as benefit receipt and benefit income.

[6] People who left work recently are counted as employed in SHARE (they were 'working in the past month') but not in EHIS (which is about current employment status). If people without disability are more likely to have *recently* left work (because periods of worklessness are shorter on average), then this may explain why the employment rate for people without disability is generally higher in SHARE.

[7] Berthoud therefore refers to this measure as 'the number of people prevented from working due to disability'. However, this is potentially confusing because we know that the disability employment gap does not capture the causal impact of disability on employment (Jones and Wass, 2013; Jones, 2006). Nevertheless, the disability employment gap is a valuable input to policymaking despite suffering from the same problem, and we therefore sometimes refer to the prevalence-adjusted measure as showing 'the percentage of the population who are potentially prevented from working due to disability'. Mostly, though, we refer to this measure as the 'prevalence-adjusted disability employment gap', a term also used in the UK Disability Employment Charter, https://www.disabilityemploymentcharter.org/faq.

[8] While calculating the prevalence-adjusted gap is straightforward, there is no simple way of constructing a confidence interval for it. Instead, we have to 'bootstrap' the confidence intervals. That is, we treat the existing sample as if it were a population and randomly sample from it to create 100 hypothetical samples, which are called 'replications'. In each of these, some sample members will appear once, some will appear multiple times, and some will not appear at all. In the simplest case, the 95% confidence interval comes from the middle 95 of these 100 replications. In fact, we can get slightly more accurate results than this with some statistical refinements, but this is still a good intuitive explanation of what is going on (see

Appendix A3 for further details). For ease of comparison, in the remainder of this chapter we also use bootstrap confidence intervals for the disability employment gap in similar fashion.

[9] DWP & DH (2017), *Work, health and disability green paper: improving lives*, https://www.gov.uk/government/consultations/work-health-and-disability-improving-lives/work-health-and-disability-green-paper-improving-lives; DWP & DH (2017), *Improving Lives: The Future of Work, Health and Disability,* https://www.gov.uk/government/publications/improving-lives-the-future-of-work-health-and-disability.

[10] https://www.gov.uk/government/news/government-hits-goal-to-see-a-million-more-disabled-people-in-work

[11] See e.g. the testimony of four academics at https://www.disabilitynewsservice.com/mps-and-experts-rubbish-governments-claims-it-cut-disability-employment-gap/ and the Parliamentary Select Committee report at https://publications.parliament.uk/pa/cm5802/cmselect/cmworpen/189/18902.htm.

[12] Raw data taken from https://vizhub.healthdata.org/gbd-results/. We have focused on Years Lived with Disability for the United Kingdom, which show a small (1%) improvement for 15-49-year-olds, and a negligible one for 50-69-year-olds.

[13] See for example 'Our Performance' in https://www.gov.uk/government/publications/dwp-annual-report-and-accounts-2022-to-2023/dwp-annual-report-and-accounts-2022-to-2023#performance-analysis, and Figure 5 in https://www.gov.uk/government/statistics/the-employment-of-disabled-people-2023/employment-of-disabled-people-2023#measures

[14] E.g. the Minister for Disabled People in 2022 said, *"The disability employment gap has closed by about five percentage points since 2013. And yesterday, the latest labour market statistics showed that we have smashed the commitment we made in our 2017 manifesto to see one million more disabled people move into employment over ten years. The fact this has been achieved now, in just half the time, reflects our…success in…getting people into work and the extra support we have introduced to help disabled people move into jobs, as well as broader changes in society, and in the workplace"* (see https://www.gov.uk/government/speeches/disability-confident-jobs-fair-speech).

[15] Indeed, the issues of methodological quality and choice of measure may inter-relate: the disability-adjusted gap may produce even more pronounced differences when comparing countries or trends over time in less-comparable surveys, where more of the cross-national differences in disability will be due to reporting and methodological effects (see Geiger *et al.*, 2017:4).

[16] This includes an open letter from one team of researchers, part of a total of 12,000 comments received by the US Census Bureau that ultimately caused their initial decision to be reversed (Cohen *et al.*, 2007; Jacobs, 2024).

[17] This echoes one of the largest international efforts to produce comparable disability statistics, the UN Washington Group on Disability Statistics (Altman, 2016; Madans *et al.*, 2011). However, there are controversies about the survey questions the Washington Group have designed to accompany this (see below), and some areas where we differ with the Washington Group on how to implement this approach. To avoid confusion, we avoid describing the impairments and activity limitation approach here as 'the Washington Group approach'.

[18] The ICF makes clear that Activity Limitations are affected by environments, but among those involved in developing the ICF there was no consensus on the relationship between Activity Limitations and Participation Restrictions (Bickenbach, 2019). The ICF manual says that *"every action, particularly when executed in a social environment, may be considered participation, and participation always entails the execution of an action or task"*, but then confusingly continues, *"Despite this relationship, the definitions of activities and participation are clearly different and distinguishing activities and participation will require careful consideration"* (WHO, 2013). Equally confusingly, the ICF beginners' guide gives users four different options for how to think about the relationship between Activities and Participation (WHO, 2002), while the ICF coding structure itself simply refers to an overall group of 'Activities and Participation' with a number of subcategories.

[19] This argument is not common in international academic debates, although see Biermann and Pfahl (2021). However, it emerged more strongly in the recent US debates, where e.g. Sabariego *et al.* (2016) criticise the Washington Group questions on the grounds that *"they claim to give a prevalence of disability but, in fact, they admit that they give a prevalence of people who are at risk of having a disability"* (emphasis added).

[20] Few people clarify this distinction, which has led to some confusion. For example, Kayess and French (2008:21) note the *"contemporary conceptual confusion between impairment and disability"*, while Lawson and Beckett (2021:21) note that the term 'disability' is sometimes used to refer to the *"social model sense of societally-created oppression or disadvantage and sometimes [it is used] to mean 'impairment'"*. Indeed, there is arguably an inconsistency between the UN Convention on the Rights of Persons with Disabilities (which focuses on potential disadvantage) and the ICF (which mostly focuses on actual disadvantage) (Bickenbach, 2019:66).

[21] To take two examples, in the United Kingdom for over 20 years, self-reported doctor-diagnosed asthma rose sharply whilst respiratory symptomology and mortality declined (Geiger, 2020); and it has been influentially noted that self-reported conditions are more common in the US than in the Indian state of Kerala, despite better objective health there (Murray and Chen, 1992).

[22] The term 'energy-limiting conditions' is preferred to the more commonly used 'fatigue' or 'exhaustion' because it better differentiates these impairments from more universal sensations of tiredness – see https://chronicillnessinclusion.org.uk/2021/04/28/what-are-energy-impairment-and-elci

[23] Rather than asking about activity limitations like the rest of the survey, instead the questions ask about the frequency and intensity of these feelings (arguably recognising that it is not always possible to build up participation restrictions from activity limitations); see https://www.washingtongroup-disability.com/resources/frequently-asked-questions/why-do-the-washington-group-anxiety-and-depression-questions-take-a-different-form-than-questions-in-other-domains/. The Washington Group's Mental Health and Psychosocial Disability subgroup is in the process of developing and testing some new functioning-based questions in these domains.

[24] Many of these criticisms argue that the individual Washington Group questions are not good at capturing the thing they are meant to capture, e.g. sensory or mobility limitations (Goddard and Hall, 2025; Hall *et al.*, 2022; Landes *et al.*, 2025; Landes *et al.*, 2024). However, it should be noted that the Washington Group team contest whether these comparisons are valid (Madans *et al.*, 2025).

[25] The DSQ is explicitly based on the social model and went through a detailed process of development. The DSQ is slightly longer than the Washington Group Extended Set (39 vs. 34 questions), and is in many

ways similar, though it deliberately omits questions on activity limitations. The DSQ takes a different approach to the Washington Group in asking about mental health and pain (it does not ask about energy-limiting impairments), asking about the frequency of these limiting people's usual activities, and then how severely this affects them. See the *Canadian Survey on Disability, 2022: Concepts and Methods Guide* (Pianosi et al., 2023) at https://www150.statcan.gc.ca/n1/pub/89-654-x/89-654-x2023004-eng.htm

[26] Note that other OECD countries are also available in the global aging data: Chile, Costa Rica, Ireland, Japan, Korea, Mexico (and also China and India, of the non-OECD countries) – see https://g2aging.org/survey-overview.

[27] Even though England/United States do not have comparable measures of self-reported activity-limiting disability, they do include the detailed impairment measures used for the disability scale. The weights for these variables are calculated using the SHARE countries, and then these weighs are used to create predicted disability scores in ELSA/HRS.

[28] While mental health has been used in previous HRS-SHARE comparisons (Jadhav and Weir, 2017; Riumallo-Herl *et al.*, 2014), the comparability of these mental health measures between surveys is limited (Courtin *et al.*, 2015). However, in our case, we find that the overall prevalence of disability is higher in ELSA/ HRS than in SHARE, but this cannot be explained by the mental health scale (which has similar prevalence in ELSA/HRS as in SHARE).

[29] The disability score uses the empirical Bayes posterior estimate of the latent variable. Note that we make two changes to the approach used by e.g. the NBER team: (1) some researchers appear to use latent variable models that are not appropriate for binary and ordinal data, but we account for this using hybrid IRT models; (2) we use a more flexible version of IRT models compared to e.g. the one-parameter rating scale models used in the World Report on Disability (WHO and World Bank, 2011), instead using two-parameter models that allow indicators to differ in both discrimination (how far they distinguish people with vs. without disability) and their severity. That said, sensitivity analyses suggest that these different latent variable methods produce similar binary disability variables, so while these choices are conceptually preferable, they are unlikely to materially affect the results.

[30] Necessarily we control for the main effects of these variables, not their interaction with each impairment measure (because to overcome reporting differences in single-item activity-limiting disability between countries, we fix the impairment weights to be constant across countries). For some indicators, we alter the form of the health measures slightly in the predicted disability models. This is because there can be collinearity in the regression model that makes some weights unexpectedly negative (Lee *et al.*, 2013), which here particularly occurred for the ADL/IADL measures. We therefore combined these into measures of 'any ADL' (binary) and 'any IADL' (0/1/2+), after which all measures have positive weights ($p<0.05$) ranging from 0.25-1.10 on the logit scale (see Appendix Table B2).

[31] One disadvantage is that it is more challenging to explain a 'coefficient on the relationship between continuous disability and employment' than a 'disability employment gap'. The other disadvantage is that there is reason to believe that the relationship between continuous disability measures and employment will be non-linear (Geiger *et al.*, 2019). It is possible to use non-linear forms of the continuous disability scale, but this then makes interpretation substantially more complex. One of the advantages of the probabilistic binary measure is that it makes no assumptions about the functional form of the relationship between disability and employment.

[32] This approach may be appealing to readers who think that the considerable cross-country differences here may still partly reflect reporting differences. We believe that these differences are reasonable given cross-country differences shown in, e.g., the WHO's Global Burden of Disease project (GBD) which estimates that in 2015, looking at the years lived with disability among 50-59 and 60-69 year olds per 1,000 people, the disability rate was 139 and 176 in the USA vs. 111 and 136 in Greece (WHO, 2016) – a 26-29% raised level in the US vs. Greece, compared to a 32% raised level for the prevalence of probabilistic disability shown in Figure 6.

[33] The Washington Group scoring is slightly more complex than this, because it is based on ordinal questions, and the cut-off is usually whether people report 'a lot of difficulty' (see also above) – but the basic intuition shown in the table holds, in that someone can be classified as 'disabled' if they answer 'a lot of difficulty' to a single question.

[34] We use the WHO (2016) summary tables for Years Lost due to Disability (YLD) for 50-59 and 60-69 year olds (which we then simple average to get a benchmark for 50-69 year olds). The country-level correlation of YLD with fixed-threshold disability is 0.39, vs. 0.47 for the probabilistic threshold.

[35] As a simple measure, the coefficient of variation (the standard deviation divided by the average) for disability prevalence in these 15 countries declines from 26% to 10% when using probabilistic vs. single-item disability.

[36] A related approach would be to have (i) single-item activity-limiting disability measures (e.g. GALI) included in all major government surveys; (ii) intermediate-length batteries of questions on impairments/activity limitations (e.g. the FDD11 version of the WHO Model Disability Survey or the Washington Group Enhanced Short Set) included in annual health surveys; (iii) full batteries of questions (e.g. the full WHO Model Disability Survey or Washington Group Extended Set) included in periodic calibration exercises.