UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

An interpretable recommendation model for psychometric data in multilabel classification and label ranking tasks, with application to gerontological primary care

Andre Paulino de Lima

Doctoral Thesis of the Postgraduate Program in Computer Science and Computational Mathematics (PPG-CCMC)

ICMC USP SÃO CARLOS

Andre Paulino de Lima

An interpretable recommendation model for psychometric data in multilabel classification and label ranking tasks, with application to gerontological primary care

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. EXAMINATION BOARD PRESENTATION COPY

Concentration area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Marcelo Garcia Manzato

USP – São Carlos

January 2026

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

L732a

Lima, Andre Paulino de
An interpretable recommendation model for psychometric data in multilabel classification and label ranking tasks, with application to gerontological primary care / Andre Paulino de Lima; orientador Marcelo Garcia Manzato. -- São Carlos, 2026.
192 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2026.

1. Recommender systems. 2. Psychometric data. 3. Multilabel ranking. 4. Interpretability and transparency. 5. Gerontological primary care. I. Manzato, Marcelo Garcia, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Andre Paulino de Lima

Um modelo de recomendação interpretável que explora dados psicométricos em cenários de classificação multirrótulos e de ranqueamento de rótulos, com aplicação em cuidados primários em Gerontologia

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. EXEMPLAR DE DEFESA

Área de concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Marcelo Garcia Manzato

USP – São Carlos

Janeiro de 2026

Para minha mãe, que me ensinou a desenhar os sons.
Para meu pai, que me ensinou a dar troco.

ACKNOWLEDGEMENTS

No começo da minha jornada no doutorado, fui convidado a colaborar com um projeto de pesquisa que previa o desenvolvimento de um sistema de recomendação de atividades relacionadas ao envelhecimento ativo para idosos. Como todo projeto em fase embrionária, apenas as ideias pilares estavam postas naquele momento, mas tínhamos uma noção clara das direções que o projeto deveria percorrer. Duas dessas ideias eram o uso de instrumentos padronizados na avaliação do potencial usuário e a exploração de atividades abertas ao público em geral. Na primeira iteração em que participei, usamos o PAGe (Plano de Atenção Gerontológica) como instrumento de avaliação e o Guia 60+ de São Carlos como catálogo de atividades. Um dos objetivos daquela iteração era coleta de dados, e isso demandava a criação de uma aplicação web para preenchimento do PAGe.

Curiosamente, o PAGe inclui em uma de suas seções finais um gráfico de radar que sumariza a avaliação do paciente. Meu parceiro de pesquisa à época, Laurentino, que ficou encarregado de desenvolver o protótipo do PAGe Online, me disse que a Brunela, a pós-doutoranda que liderava aquele esforço e é uma gerontóloga com experiência prática, tinha dito que o layout do formulário poderia ser adaptado da forma que fosse necessária no PAGe Online, mas o diagrama ao final da avaliação era inegociável. Esse comentário me chamou a atenção. De que forma esse diagrama é usado pelo gerontólogo que acabou de conduzir a avaliação? Ou ele seria útil a outros profissionais, em outros momentos? Ele serve de apoio na construção do plano de cuidados de alguma forma? As perguntas se multiplicaram. Um semestre antes, eu tinha feito a análise fatorial de um dataset da área de sistemas de recomendação como parte da disciplina de IHC. As ideias foram se misturando, o vínculo entre modelo congenérico em psicometria e representação computacional começou a ganhar contornos. Não levou muito tempo para perceber que a própria representação gráfica, preferida pelos profissionais, poderia ser empregada como representação interna do sistema de recomendação. O resto dessa história está escrito nos capítulos desta tese.

Gostaria de começar meus agradecimentos aos integrantes do meu grupo de pesquisa. Primeiro, ao meu orientador, Marcelo Manzato, cuja paciência eu pus à prova em inúmeras ocasiões. Obrigado pela autonomia, pelo incentivo e pela compreensão nos momentos em que a vida pessoal impôs atenção exclusiva. À Maria da Graça Pimentel, pelo convite para colaborar com o projeto que estava sendo desenvolvido no âmbito do ESPIM e por nos apresentar aos grupos de pesquisa que se tornaram mais tarde parceiros nesse projeto. Ao Laurentino Dantas pelo companheirismo e pelas discussões criativas, críticas e divertidas. Todas foram úteis e necessárias. À Brunela Orlandi, por atuar como nossa consultora em gerontologia, mesmo depois da conclusão do pós-doutorado. Em especial pela ajuda na navegação da literatura especializada e nosso grupo de discussão sobre “idade biológica”.

Sendo um engenheiro de computação, atuar em um projeto de pesquisa aplicado na área da saúde exigiu algumas adaptações. Começo agradecendo à Professora Ruth Melo da Escola de Artes, Ciências e Humanidades da USP por ter atendido ao pedido que eu e o Laurentino fizemos para assistir a disciplina de Avaliação Gerontológica Ampla como ouvintes. Estendo o agradecimento à turma da Gerontologia que cursou essa disciplina no segundo semestre de 2021, pela recepção calorosa e pela participação nos estudos com usuários que Laurentino e eu conduzimos naquele semestre. Agradeço também ao Professor Julio de Rose do Departamento de Psicologia da UFSCar por ter permitido que eu assistisse a disciplina de Análise Comportamental da Cognição como ouvinte, e também aos colegas da turma pela recepção atenciosa e amigável.

Desenvolver pesquisa em sistemas de recomendação na área da saúde foi um desafio que eu não teria superado sem o apoio dessas pessoas: Paula Castro, do Departamento de Gerontologia da UFSCar, Ruth Melo, Rosa Marcucci e Suzana Andrade, afiliadas/ligadas ao grupo de Gerontologia da Escola de Artes, Ciências e Humanidades da USP. Obrigado por terem compartilhado comigo amostras de conjunto de dados que vocês coletaram. A amostra de dados de idosos saudáveis avaliados com o instrumento WHOQOL-BREF, assim como a amostra de dados de pacientes avaliados com o instrumento AMPI-AB foram fundamentais não só para a elaboração da estratégia de avaliação do modelo que proponho nesta tese, mas também para fundamentar minha compreensão de como diferentes instrumentos capturam variação em populações de idosos.

Por fim, mas não menos importante, agradeço ao apoio institucional. Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"Unwarranted variation in health care delivery is the variation that cannot be explained on the basis of illness, medical evidence, or patient preference — is ubiquitous."

John Wennberg

(on the impact of unwarranted variation in care on people's lives)

"I feel passionately about measurement — about how difficult it is, about how much theory and conceptualisation is involved in measurement, and indeed, how much politics is involved."

Angus Deaton

(on the role of measurement in science and policymaking)

"... for geometry, you know, is the gate of science, and the gate is so low and small that one can only enter it as a little child."

William Clifford

ABSTRACT

LIMA, A.P. **An interpretable recommendation model for psychometric data in multilabel classification and label ranking tasks, with application to gerontological primary care**. 2026. 192 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

Recommender systems are technology components that typically use behaviour data to learn a person's preference for an item. They have been applied to diverse domains with astounding success, but there are challenges that must be overcome to make them useful in healthcare settings. The reasons are varied: the lack of publicly available clinical data, the difficulty that users may have in understanding the reasons why a recommendation was made, the risks that may be involved in following that recommendation, and the uncertainty about its effectiveness. In this monograph, these challenges are addressed by the provision of visual explanations that are faithful to the recommendation model and interpretable by care professionals that use psychometric instruments to assess their patients. We propose a recommendation model that leverages the structure of psychometric data to produce recommendations based on a dataset with patient assessments and health interventions. The discussion is centred around gerontological primary care, a niche where professionals rely on psychometric instruments to comprehensively assess their patients. We applied the model to this niche to illustrate how it can assist the attending professional in the creation of personalised care plans. We report the results of a comparative offline performance evaluation of proposed model on healthcare datasets that were collected by local research partners in Brazil. We also report the results of a user study that evaluates the interpretability of the visual explanations the model generates. The results suggest that the proposed model can promote the application of recommender systems in this healthcare niche, which is expected to grow in demand, opportunities, and information technology needs as demographic changes become more pronounced.

Keywords: Recommender systems. Psychometric data. Multilabel classification. Label Ranking. Interpretability. Visualisation. Gerontological primary care.

RESUMO

LIMA, A.P. **Um modelo de recomendação interpretável que explora dados psicométricos em cenários de classificação multirrótulos e de ranqueamento de rótulos, com aplicação em cuidados primários em Gerontologia**. 2026. 192 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

Os sistemas de recomendação são componentes tecnológicos que normalmente utilizam dados comportamentais para aprender a preferência de uma pessoa em relação a algum item. Eles têm sido aplicados a diversos domínios com um sucesso surpreendente, mas existem desafios que precisam ser superados para torná-los úteis em aplicações na área da saúde. Os motivos variam: a falta de dados clínicos disponíveis publicamente, a dificuldade que os usuários podem ter em entender os motivos pelos quais uma recomendação foi feita, os riscos que podem estar envolvidos em seguir essa recomendação e a incerteza sobre sua eficácia. Nesta monografia, esses desafios são abordados por meio do provimento de explicações visuais que são fiéis ao modelo de recomendação e interpretáveis por profissionais de saúde que utilizam instrumentos psicométricos na avaliação de seus pacientes. Propomos um modelo de recomendação que explora a estrutura dos dados psicométricos para produzir recomendações com base em um conjunto de dados com avaliações de pacientes e intervenções de saúde. A discussão é focada em cuidados gerontológicos na atenção primária, um nicho da saúde no qual os profissionais utilizam instrumentos padronizados para avaliar de forma abrangente o estado de saúde dos pacientes. Aplicamos o modelo neste nicho para ilustrar como ele pode auxiliar o profissional de saúde na criação de planos de cuidados personalizados. Os resultados de uma avaliação comparativa offline do desempenho do modelo proposto são reportados. Essa avaliação emprega uma coleção de datasets da área da saúde que foram coletados por parceiros de pesquisa locais no Brasil. Também são reportados os resultados de um estudo com usuários que avalia a interpretabilidade das explicações visuais geradas pelo modelo. Os resultados sugerem que o modelo proposto pode promover a aplicação de sistemas de recomendação nesse nicho da saúde, que deve apresentar crescimento em demanda, oportunidades e necessidade de tecnologia da informação à medida que as mudanças demográficas se tornarem mais acentuadas.

Palavras-chave: Sistemas de recomendação. Dados psicométricos. Classificação multirrótulo. Ranqueamento de rótulos. Cuidados primários em Gerontologia.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AMB	Brazilian Medical Association (Associação Médica Brasileira)
ADL	Activities of Daily Living
AMPI-AB	Multidimensional Evaluation of Older People in Primary Care (Avaliação Multidimensional da Pessoa Idosa na Atenção Básica)
CGA	Comprehensive Geriatric/Gerontological Assessment
CFA	Confirmatory Factor Analysis
ELSI	Brazilian Longitudinal Study of Aging (Estudo Longitudinal da Saúde do Idoso - Brasil)
Fiocruz	Oswaldo Cruz Foundation (Fundação Oswaldo Cruz)
FESC	Educational Foundation of São Carlos (Fundação Educacional São Carlos)
GSA	Gerontological Society of America
IADL	Instrumental Activities of Daily Living
IC	Intrinsic Capacity
ICOPE	Integrated Care for Older People
NIH	National Institutes of Health (US)
PAI	Elderly Caregiver Program (Programa Acompanhante de Idosos)
PELI-NH	Preferences for Everyday Living Inventory for Nursing Homes
SBGG	Brazilian Society of Geriatrics and Gerontology (Sociedade Brasileira de Geriatria e Gerontologia)
SUS	Brazilian National Health System (Sistema Único de Saúde do Brasil)
U3A	University of the Third Age (Universidade da Terceira Idade)
WHO	World Health Organisation, the United Nations agency that works to promote health, keep the world safe, and serve the vulnerable.
WHOQOL	WHO's Quality of Life assessment instrument

LIST OF SYMBOLS

A	A generic matrix (uppercase latin letter). We refer to the i-th row of A as A_i (a row vector), $A_{:j}$ denotes its j-th column (a column vector), and a_{ij} is an element of A. Unless stated otherwise, A is a real matrix.
$\mathring{A}, \hat{A}$	$\mathring{A}$ denotes a matrix that holds data in their original form (as opposed to transformed data), and $\hat{A}$ denotes a reconstructed version of A.
a	A generic variable (lowercase latin letter). Whenever no ambiguity arises, we may use a to represent a (row or column) vector from the matrix A.
m	The number of assessments in a dataset, indexed by $i \in 0 \dots (m-1)$
n	The number of labels in the dataset, indexed by $j \in 0 \dots (n-1)$
d	The number of scores per assessment, indexed by $k \in 0 \dots (d-1)$
n_a	The number of annuli in a Polygrid diagram, with $p \in 0 \dots (n_a-1)$
n_s	The number of sectors in a Polygrid diagram, with $q \in 0 \dots (n_s-1)$
n_{as}	The number of annular sectors, indexed by $r \in 0 \dots (n_{as}-1)$
X	The (m, d)-matrix of assessment scores. Sometimes, we refer to the data of two arbitrary subjects, Alice and Bob, as X_a and X_b.
Y	The (m, n)-matrix of assignment data (label presence or label ranking)
U	An (m, n)-matrix with u_{ij} being the degree of membership of the subject X_i to the fuzzy cluster identified with the j-th label of Y
$\overline{\mathbb{D}}$	The closed unit disc centred at the origin of the complex plane $\mathbb{C}$
ζ	The enumeration of the d-th roots of unit $(\zeta_0, \dots, \zeta_{d-1}) : \zeta_k^d = 1,\ \zeta_k \in \partial\overline{\mathbb{D}}$
Ω	A partitioning of $\overline{\mathbb{D}}$ into disjoint annular sectors $(\omega_0, \dots, \omega_{n_{as}-1})$
Z	The (m, d)-matrix of polygons on the unit disc $\overline{\mathbb{D}}$ (a complex matrix)
S	The (m, n_{as})-matrix with the decomposition of the polygons in Z
W	The (n, n_{as})-matrix with the weights given to each pair of label and cell
$\mu(\dots)$	A function that "measures" the area of an arbitrary planar shape in $\overline{\mathbb{D}}$
$\triangle(\dots)$	A mark to remind the reader that the tuple of complex numbers at its right was built so that it specifies a simple, solid polygon

CONTENTS

1 INTRODUCTION

Recommender systems are technology components that employ human behaviour data to perform predictive tasks. When applied to the ever-growing e-commerce domain, a recommender system provides recommendations based on consumer behaviour data, such as who visited which product web page, who bought which product, or how a product was rated or described by those who bought it. In other domains, a recommender system can supplement behaviour data with demographic, anthropometric, physiological, or environmental data to improve its performance (Stiller; Roß; Ament, 2010; Rist *et al.*, 2015; Orte *et al.*, 2018). Like any other machine learning application, recommender systems critically depend on quality data, but do not perform data collection. So where do such data come from? Actually, in many commercial deployments, recommender systems are modules within a larger software platform that provides the supporting services required by the recommender, such as collecting and preparing data, and managing the user interface.

Regarding the predictive task they perform, the general goal is to predict preference behaviour in an opportunistic way, so that the system can produce recommendations that are relevant and timely. Building on the e-commerce example, the system may start recommending specific brands of computer accessories based on the fact that the user has recently purchased a notebook or has visited multiple web pages with notebook offers. In more complex settings though, the goal may be to predict whether the outcome of an action would benefit the user. Take for instance the application reported by Gannod *et al.* (2019). The Preferences for Everyday Living Inventory - Nursing Homes (PELI-NH) is a questionnaire that captures preferences of nursing home residents regarding 72 aspects of care provided by these institutions, which can be tailored to meet specific needs or tastes. The length of the questionnaire is, however, often seen as a barrier to its use. To address this issue, the reported system combines the resident's answers given to a shorter version of the questionnaire (16 questions) with a dataset of hundreds of fully-answered questionnaires. Then, it recommends to the nursing home manager which additional preferences from the PELI-NH inventory the resident would probably benefit from changing the default choice.

The two examples described above illustrate how recommender systems can differ markedly in the type of data they use and the predictive tasks they perform depending on their domain of application. At least in part, these differences explain why applications of recommender systems in healthcare are relatively rare when compared to those in more traditional domains (Croon *et al.*, 2021). The amount of resources needed to gather data and to demonstrate the value of the system to its potential users is expected to be larger for applications in healthcare. This is due to barriers to development, acceptance, and deployment that seem to a large extent specific to the healthcare domain, such as:

- Public datasets play a critical role in the research and development of recommender systems. The collection of healthcare data is expensive and must be conducted in compliance with local regulations, such as HIPAA in the US, GDPR in the EU and LGPD in Brazil. These regulations also govern with whom the data can be shared, which limits the research groups that will have access to the datasets.

- The acceptance of a recommender system as a support tool by a specialised healthcare community, such as geriatric care professionals, depends on several factors, including the system's ability to provide convincing explanations for recommendations made (Nunes; Jannach, 2017). Despite recent efforts to enhance this ability in diverse settings, further research is badly needed (Mamalakis *et al.*, 2024).

- The deployment of an AI-based system as a support tool in a clinical setting is a complex process, which involves reporting results according to guidelines that are still in development (Collins *et al.*, 2024). Moreover, recent events have drawn worldwide attention to the risks of unchecked applications of AI in public affairs (Bradshaw; Howard, 2018; Watson, 2022). Responses to these (and similar) events may harden barriers to recommender systems in high-stakes domains (Gilbert, 2024).

These differences grow deeper under a more careful analysis. The community's effort has been rightfully directed to develop recommender systems that perform well in low-stakes domains where voluminous, sparse data are publicly available for evaluating new recommendation models. Typical benchmark datasets, like many found in the GroupLens and ACM RecSys repositories[1], contain millions of data points, each representing the manifest/implicit preference of a user regarding some item. Although these datasets hold data about the preference of thousands of users regarding thousands of items, each user rates only a small and distinct fraction of all available items. Moreover, operating in a low-stakes domain implies that (a) any consequences of making irrelevant recommendations are limited to the users' attitudes towards the system itself (e.g., decreased trust in the system's ability to provide good recommendations) and (b) the ability to provide explanations is probably not a critical success factor for the application (Tintarev; Masthoff, 2022).

In contrast, some niches in healthcare may be better described as a high-stakes, dense-data domains. Owing to professional ethics in healthcare, patients must be thoroughly assessed using instruments whose effectiveness has been confirmed, and receive the care they need accordingly. For instance, geriatric and gerontological practices in many countries, including Brazil, encourage the use of standardised instruments (e.g., questionnaires) to guide the comprehensive assessment of patients in primary care, and many professional associations promote their own inventories (Brown *et al.*, 1988; Gorzoni, 2017). This policy has important consequences for devising recommender systems to operate in this niche.

[1] These repositories can be found here (Grouplens) and here (ACM RecSys).

First, every patient is assessed along health dimensions defined by the instrument. Thus, if the instrument includes an assessment of cognitive capacity, then a score of cognitive capacity will be available for every patient in a dataset. This is analogous to having a dataset of movie ratings in which every user has rated exactly the same set of movies. Moreover, the ratings would have been collected in a more systematic manner, for example, by having each rater answer a questionnaire just after watching the movie in a movie theatre. In other words, different from the movie recommendation domain, in which the data describing the users' spontaneous rating behaviour are heavily sparse, the data are dense and structured in this niche, consisting of scores a patient obtains along a set of health dimensions, and the scores are computed following a standardised procedure.

Second, the number of items to be recommended is minute compared to that of more traditional applications of recommender systems. For example, in a recent study, healthcare researchers collected the assessment of intrinsic capacity[2] of more than ten thousand older participants (60+ years) and, after a screening process, the recommendations and referrals[3] made for each remaining participant were recorded. In total, 22 distinct recommendations or referrals were made by the attending primary care professionals (Tavassoli *et al.*, 2022, see Table 3). In this setting, a recommender system could be created to advise the care professional about which "items" should be considered, based on the outcome of a patient's assessment of intrinsic capacity. The predictive task performed in this case can be formally defined as a multilabel classification or as a label ranking task, as will be detailed later.

Last but not least, there is the need to minimise risks of harmful recommendations. The obvious and probably the safest solution is to adopt an expert-in-the-loop approach, as was done in the nursing home example given earlier (Gannod *et al.*, 2019). Unlike the usual practice in low-stakes domains, where the final user is faced with the recommendation and appraises its relevance, in an expert-in-the-loop approach, the recommendation is presented to the care professional, who exerts her judgement about its adequacy and decides whether the recommendation should be taken into account when creating the patient's care plan. To assess whether a recommendation is adequate, the expert would probably benefit from an explanation about the recommendation made (Nunes; Jannach, 2017). However, any provided explanation should preferably be (a) faithful — in the sense that it reflects what the recommendation model actually computes, and (b) easily interpretable by the experts interacting with the system.

---

[2] Intrinsic capacity (IC) is defined by the WHO as "the combination of the individual's physical and mental, including psychological, capacities" that are essential to her everyday functioning. An instrument to assess IC evaluates the individual's capacity along five dimensions: cognitive, psychological, sensorial, locomotive, and vitality.

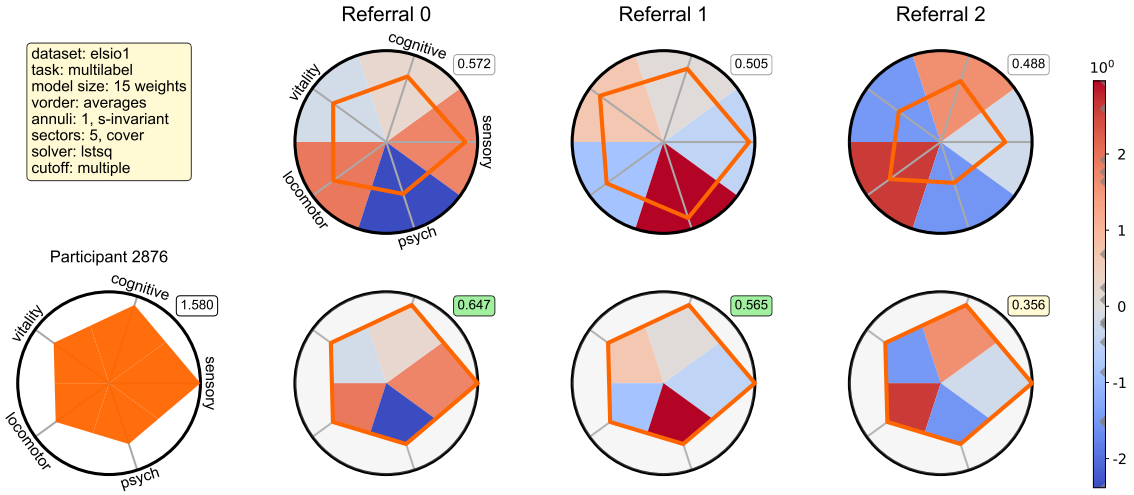
[3] In healthcare, a referral is the process by which a healthcare provider directs a patient to another healthcare professional or specialist for further evaluation, diagnosis, or treatment. This typically occurs when the referring provider believes that the patient's condition requires expertise or services beyond their scope of practice.

In summary, devising recommender systems for healthcare requires overcoming barriers that are mostly absent from the lifecycle of more traditional applications. We believe that any research strategy to be successful in this challenge needs to recognise that each healthcare niche likely has its relevant data structured differently, and to acknowledge that opportunities to create impact will come from investigating decision-making tasks that are valued by specialists in these niches. Acting on this belief, we propose a new recommendation model that focuses on a narrow niche, gerontological primary care, with the aim of leveraging the structural characteristics of the psychometric instruments used by professionals to assess their patients. More precisely, the proposed recommendation model takes advantage of the fact that the psychometric data describing patients are dense, low-dimensional, and meaningful to produce recommendations and explanations. The explanations follow a visual style and correspond to an interactive diagram. Furthermore, these explanations are faithful in that they depict what the recommendation model actually computes, and are scrutable because the user can inspect the values ascribed to distinct components of the diagram. The predictive task performed by the recommendation model is to suggest health interventions based on the patient's assessment data. This means that, in an expert-in-the-loop use case, the recommender system advises the attending primary care professional about which interventions should be considered beneficial for the patient.

To give the reader a more concrete grasp of the proposed explanation style, we present an example in Figure 1. Building on the study of intrinsic capacity and referrals mentioned earlier (Tavassoli *et al.*, 2022), the diagram shows an explanation of why some (fictitious) referrals are recommended for a (real) person. The diagram is composed of a number of radar charts disposed on a grid: (a) the outcome of the patient's assessment is represented by the radar chart in the first column, (b) the referrals are represented in the first row, and (c) the matching between the patient assessment and a referral is shown in a separate chart in the second row, right below the chart of the corresponding referral.

The idea of representing the patient's assessment as a radar chart comes from the literature on geriatrics and gerontology. As will be reviewed in Section 2.2, several authors have highlighted over the years the practical advantages both for professionals and patients of using this diagram to represent the results of a comprehensive assessment. We built on this idea by giving the radar chart a footing on psychometric models. Specifically, we demonstrate a relationship between measurement and area: the position of an individual on the measurand (i.e., the latent variable being measured by the psychometric instrument) has a monotonic relationship with the area of the assessment polygon. In Figure 1, the measurand is intrinsic capacity, and the vertices of the assessment polygon are the scores the patient obtained for the sensory, cognitive, vitality, locomotor, and psychological domains. The numeric value of the area of the polygon is shown in a tag that appears next to the chart. The measurement-area relationship implies that the higher the intrinsic capacity of an individual, the larger the area of the polygon representing their assessment.

Figure 1 – An example of the proposed explanation style



Source: The author.

Legend: The explanation diagram is composed of a number of radar charts disposed on a rectangular grid. There are three types of charts: (a) assignment charts, which are placed on the first row, (b) assessment charts, on the first column, and (c) the matching charts.

Regarding the representation of referrals (or interventions, more generally), each radar chart contains three elements: a polygon, the colours that fill the disc, and a tag. The vertices of the polygon are the average scores obtained by a group of patients who were previously assigned to that referral by experts. The colours represent weights given to different cells of the partition of the disc, and the tag shows the threshold for this referral. The colour bar on the right of the diagram maps each colour to its respective weight.

The matching between the patient's assessment and a referral corresponds to a recommendation decision. The matching charts also contain three elements: a polygon, the colours that fill it, and a tag. The polygon is always a copy of the polygon that represents the patient's assessment, and it is filled with the colours that appear in the corresponding referral. The number shown on the tag is the weighed area of the resulting polygon: if it is greater than the referral's threshold, it appears in green to indicate that the referral is recommended to the patient; otherwise, it appears in yellow. In interactive mode, the user can inspect the weight of a cell or the area of a polygon within a cell by clicking on it.

Finally, the recommendation model learns the weights that better reproduce the collective judgement of the healthcare experts who built the dataset of patient assessments and assigned interventions. Similarly to collaborative filtering techniques, both the final users ("patients") and recommended items ("referrals") are represented in the same abstract space: the unit disc. The inner product between two such representations is the value that appears in the matching chart's tag. In the next chapters, we will argue that the proposed recommendation model is transparent and interpretable, and also detail how the many variations of this explanation style are supported by the proposed model.

1.1 Motivation

The United Nations has declared the decade of 2021-2030 as the Decade of Healthy Ageing. This development builds on sustained efforts from the WHO to advance healthy ageing as the process of developing and maintaining the functional ability that enables well-being in older age[4]. These efforts are partly motivated by the fast demographic changes that are underway worldwide: "For the first time in history, most people can expect to live into their 60s and beyond." (World Health Organization, 2015). Although these changes may bring opportunities for both individuals and societies, an increase in the demand for health services is expected. National health systems will need more resources to cope with this new reality but, as argues Hausman (2015), the allocation of health-related resources relies on political deliberation. Being the latter ever more difficult to attain, alternative routes should be paved. A long-term strategy for coping with an increasing demand is to optimise the use of existing resources, which makes the referral process a promising target.

Three facts in support of the latter assertion are offered. First, a recent report by the WHO compiled lessons learned from several EU member states during the course of the COVID-19 pandemic, directed at improving referral systems worldwide (World Health Organization, 2023). These lessons show that, under an abrupt and large increase in demand, improvements to the referral process led to better outcomes both for individuals and institutions. Second, many studies in the public health literature suggest that the variation in referral rates among professionals in primary care is significant[5]. This variation in referral rates have long called the attention of researchers and policymakers, as it may imply financial burden and unfavourable outcomes for patients, and other negative impacts on institutions (O'Donnell, 2000; Mullan, 2004; Shashar *et al.*, 2023). Public health experts have long deemed this variation unwarranted because it is observed even among professionals performing similar functions in the same geographical area (e.g., general practitioners working in the same city or hospital). Although it remains largely unexplained, some evidence suggests that modifiable psychological factors associated with tolerance to uncertainty may play a role in this variation (Shashar *et al.*, 2023). Third, and of local relevance, there is evidence that the referral system implemented in SUS/Brazil also suffers from such difficulties and inefficiencies, as reported recently in the Atlas of Variation in Healthcare in Brazil (Diegoli *et al.*, 2022).

[4] Functional ability refers to these abilities (among others): autonomy, conceived as the capacity and the right to make own choices, and independence, as the ability to perform tasks of daily life without assistance, such as getting in and out of bed, using the toilet, taking medications, bathing, eating, preparing meals, and shopping for groceries. These functions allow for a person to remain socially relevant and to cope with adverse life events.

[5] In this context, referral rate is the ratio a/b, where (a) is the number of visits to a referring institution that have resulted in the patient being referred to a receiving institution, and (b) the number of patient visits to the referring institution (World Health Organization, 2023).

In summary, the referral process is a promising target because there is room for improvements, as well as incentives for pursuing them. An interpretable recommendation model that learns from a dataset of standardised patient assessments and referrals made (or reviewed) by experienced practitioners can assist students and novice professionals in developing, improving, or reflecting on their referral practices. In the long term, it may contribute to reducing unwarranted variation in referrals. More ambitiously, it may also help researchers improve existing instruments, as the model provides a way to visualise and inspect the predictive relationship between latent variables and related outcomes.

In addition, and specifically for the research community on recommender systems, this work may encourage other researchers investigate and explore representations of users and items that differ from the long-established vector representations. This may lead to new ways in which explanations of recommendations can be produced, which in turn may reduce barriers to the application of recommender systems to domains where the provision of explanations is mandatory and faithful explanations are highly preferable.

1.2 Objectives

The primary objective of this research project was to develop a recommendation model that can leverage the structure of psychometric data to produce recommendations that can be meaningfully explained to users who are familiar with the instrument. To accomplish this main objective, the following specific objectives were pursued:

- To gather datasets with psychometric data collected in gerontological primary care settings, and to extend them in principled ways to meet our research needs. This objective demanded partnering with groups that conduct gerontological research.

- To develop a recommendation model for psychometric data that produces faithful and interpretable explanations to an expert-in-the-loop. The model is described in the context of its application in gerontological primary care, as a supporting tool to assist the care professional in the creation of a personalised care plan to the patient.

- To assess the performance of the proposed model in an offline experiment, using the datasets with psychometric data that we gathered. The evaluation methodology is an adaptation of standard procedures used to assess the performance of machine learning models in multilabel classification and label ranking tasks to our setting.

- To assess the interpretability of the proposed model in a user study, using a dataset about a topic that is familiar to the general public. The evaluation methodology follows standard procedures used to assess the effectiveness of visualisations in supporting lay users in decision-making tasks.

1.3 Contributions

The main contributions of this research project can be summarised as follows:

- A recommendation model for psychometric data with faithful visual explanations. The model explores the fact that the data are low-dimensional and offers a mechanism to produce higher-dimensional representations of the data. The forward computational graph of the model is visually encoded in the explanations. Moreover, the area of the element that represents the patient's assessment in the explanation diagram has a monotonic relationship with the latent variable underlying the psychometric instrument, which grounds its interpretation as a visual analogue of the measurand.
- An intuitive visualisation that supports decision-making tasks that can be modelled as a multilabel classification or ranking task, as seen in Figure 1. The visualisation is said to be intuitive because its design nudges the participant's attention to focus on the depiction of the assessment polygon rather than on the individual numerical scores from the assessment. The advantages of this nudging are empirically shown. More generally, the visualisation can be used to display the multiplication between small matrices or a small tensor, as in Figure 18.
- A useful visualisation that supports the inspection of the model's performance. As shown in Figure 19, the "scales diagram" dissects the predictive performance of the model from a perspective afforded by the geometry of the feature space. In contrast with the previous visualisation, which shows a few assessments in a detailed way, this visualisation depicts simultaneously all instances of the dataset using the assignment weights as "basis" vectors to coordinate the feature space (or subspace).
- A conceptual framework for interpretability as a multidimensional property. The framework reduces ambiguity about its contributing factors: data (i.e., dimensionality and meaningfulness of attributes), the model's architecture (scalability and ability to preserve the meaning of input data), transparency, and the users. Transparency is conceived as a relationship between the model and the explanations it generates, and interpretability is operationalised as the outcome of an experiment.

An early version of these contributions was published in the following conference article: LIMA, A. P. de *et al.* An interpretable recommendation model for gerontological care. *In*: **Proceedings of the 15th ACM Conference on Recommender Systems**. New York, NY, USA: ACM, 2021. (RecSys '21), p. 620–626. ISBN 9781450384582. Available at: https://doi.org/10.1145/3460231.3478850.

A mature version of these contributions has been submitted to the reference journal of the community on recommender systems (TORS), and an author version is available

in the following archive: LIMA, A. P. de *et al.* **An Interpretable Recommendation Model for Psychometric Data, With an Application to Gerontological Primary Care**. 2026. Available at: https://arxiv.org/abs/2601.19824.

As secondary contributions of this project, we would like to point out that:

- We share the computer code that extracts a dataset of intrinsic capacity assessments from the "ELSI-Brazil Wave 1" dataset (Lima-Costa *et al.*, 2018) using a method described in Aliberti *et al.* (2022). Access to the dataset must be requested from the research team that maintains the original data, via ELSI-Brazil project website.
- We also share a software environment in which both the computational methods and visualisation described in this thesis can be examined. A domain-specific language was created to abstract the details involved in training, assessing, and comparing models, allowing others to explore or reproduce the results presented in the thesis.

The code just described can be found in the project's github repository (click here).

1.4 Outline

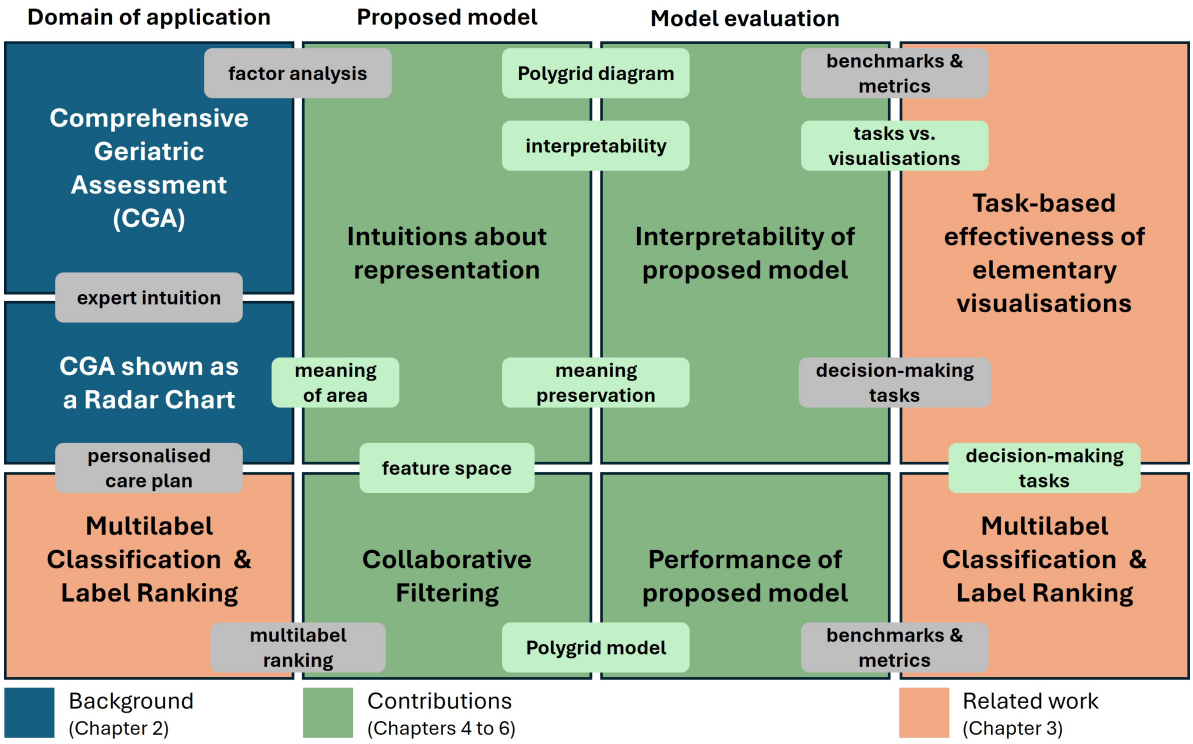
The main topics and the connecting ideas presented in this work are illustrated in Figure 2, which we use to explain how the remainder of this monograph is organised.

Chapter 2 briefly discusses how gerontologists foster the well-being of older people in primary care and describes the role of a medical procedure known as a comprehensive geriatric assessment for this purpose. This standardised assessment is usually based on a questionnaire developed in the psychometric tradition of factor analysis, which is also briefly reviewed. The chapter also reviews a recurring topic in the literature on gerontology about the use of radar charts to represent the results of a patient's assessment.

Chapter 3 begins with a review of the literature on healthcare recommender systems to clarify the gap in applications targeting older people in the context of gerontological care. The review shows that applications that use standardised assessments are relatively rare, and systems that provide faithful explanations are even rarer. This is followed by a brief review of machine learning models for multilabel classification and label ranking tasks. This topic is important because the creation of a personalised care plan for the patient in primary care can be formalised by these tasks. The chapter concludes with a brief review of the literature on visualisation research focused on visualisation-supported decision-making tasks. We approach this topic because our model generates visual explanations. The latter two reviews provide us with the conceptual and methodological tools that will be later used to organise the evaluation of performance and interpretability of the proposed model.

Chapter 4 introduces Polygrid, our proposed recommendation model. We detail the model's learning pipeline step by step, first covering multilabel classification tasks and then

Figure 2 – Outline of main topics and connecting ideas of this thesis



Source: The author.

Legend: A large box represents a topic, and its background colour indicate whether it reviews background or related work, or introduces our contributions. Grey tags represent the main ideas connecting two topics. Green tags represent the main ideas connecting topics of our original contribution.

label ranking tasks. The chapter illustrates and discusses the effects of different choices for the model's main hyperparameters, and concludes with two separate discussions. The first addresses the model's learnability, laying out a theoretical framework to explain why the model successfully learns to perform the target tasks. The second discussion focuses on the interpretability of the proposed model. Here, we attempt to navigate definitional issues in the literature by defending an operationalist view of interpretability, combined with an axiomatic framework in which we introduce a notion of meaning preservation.

This is followed by two evaluations. Chapter 5 reports the results of a comparative offline evaluation against seven models for multilabel classification and label ranking tasks on fifteen healthcare datasets. The datasets of patient assessments were collected with psychometric instruments by partner research groups, and the labels were created synthetically (with one exception). Chapter 6 reports the results of a within-subjects user study to evaluate the interpretability of the Polygrid diagram. Participants are presented with an explanation diagram and an input case, and must correctly classify the case. The Iris dataset is used to specify the decision-making tasks of the study. Finally, Chapter 7 concludes our work, highlights limitations, and presents opportunities for future work.

2 BACKGROUND

In this chapter, we will briefly review how gerontologists promote well-being of older people, and the tools they use. We describe the important role that comprehensive assessments play in the care of older people, and then show that these assessments are founded on psychometric methods. The first part will be used in the next chapter to show that healthcare recommender systems are not aligned with current gerontological practice, and the second part in a later chapter to describe the structure of psychometric data.

2.1 How gerontologists assess and foster well-being of older persons

The current notion of ageing being advanced by the WHO states that the health status and functional ability that an older person preserves are strongly influenced by her life course, and loosely associated with her chronological age (World Health Organization, 2015, pp. 29)[1]. In this view, functional ability is the person's capacity to perform the activities she finds meaningful and valuable in life. This capacity emerges from complex interactions between her personal characteristics (such as sex, gender, ethnicity, wealth) and the environment she inhabits (her home, community, and broader society). These interactions shape opportunities, barriers, exposures, and access to resources such as health services. A long-term consequence of these interactions is the wide variance in the levels of functional ability observed within and among older populations around the world.

The WHO also advances the notion of healthy ageing as "the process of developing and maintaining the functional ability that enables well-being in older age." It is important to emphasise that this notion of healthy ageing is not synonymous with a disease-free state (which is problematic in older age because many individuals develop chronic conditions that, if effectively managed, do not substantially decrease functional ability), but with a state in which individuals maintain the ability to pursue the things they value. Healthy ageing is founded on the belief that "for most older people, the maintenance of functional ability has the highest importance" (World Health Organization, 2015).

Gerontology, as a practice, is aligned with this view of healthy ageing, as professionals seek to foster functional ability in older populations so that people can experience well-being for as long as possible (Melo; Silva; Cachioni, 2015). In a consensus development conference sponsored by the NIH and other institutions in 1987, specialists in the care of older people have agreed that their Comprehensive Geriatric Assessment (CGA) procedures should be improved. These improvements should aim how specialists (a) select interventions to restore or preserve health and functional status, (b) predict health outcomes, and

[1] However, for practical reasons, a specific age is usually adopted as a reference: In Brazil, a federal law establishes that an older person is a citizen aged 60 years or over (Tebet, 2006).

(c) monitor clinical change over time (Brown *et al.*, 1988). In Brazil, CGA is currently defined as a medical procedure by the AMB (Associação Médica Brasileira), and the SBGG (Sociedade Brasileira de Geriatria e Gerontologia), the main association of geriatricians and gerontologists, promotes a standard CGA inventory to its members (Gorzoni, 2017).

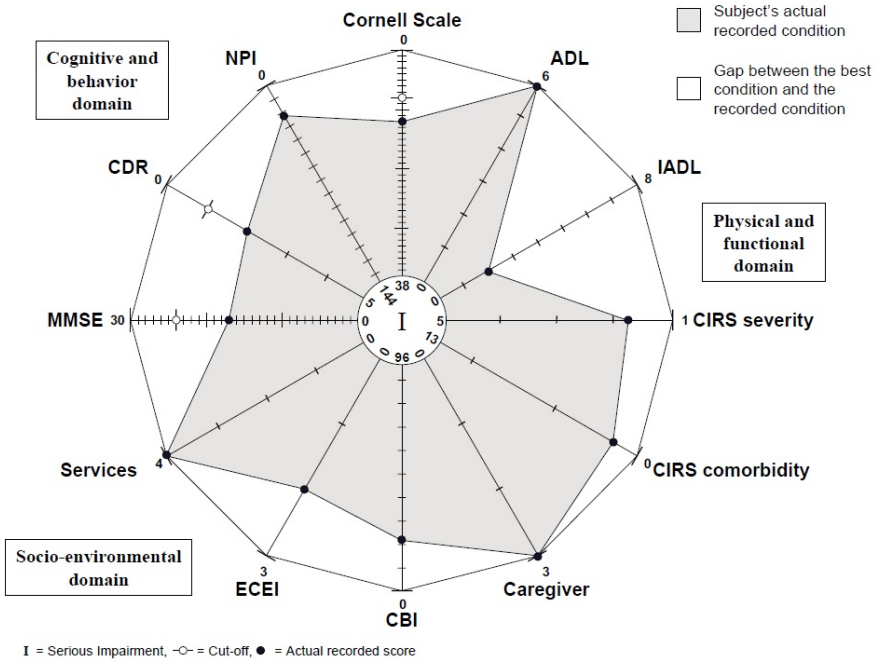
In general, CGA inventories are structured questionnaires designed to guide care professionals in assessing their geriatric patients. These inventories usually include items that survey health domains such as physical medical conditions, mental health conditions, functioning, and social and environmental circumstances, providing the full biopsychosocial nature of the individual's capacities, problems, and conditions (Welsh; Gordon; Gladman, 2014; Garrard *et al.*, 2020). Based on the patient's responses, a set of numerical scores that reflect the individual' status along the surveyed domains can usually be computed by following a standardised procedure. How these inventories and their scoring procedures are developed is the subject of Section 2.3. It must be noted that, when selecting interventions to add to the patient's care plan, the attending care professional bases her clinical judgement not only on the results of the assessment, but also on the attitudes, preferences, and expectations of the patient and their family (World Health Organization, 2017). Frequently, the latter are only captured as non-verbal cues, which may be verbally clarified by the professional. In our view, this fact adds another reason why the adoption of an expert-in-the-loop is the way forward for recommender systems in healthcare.

2.2 The visual display of the results of a patient assessment

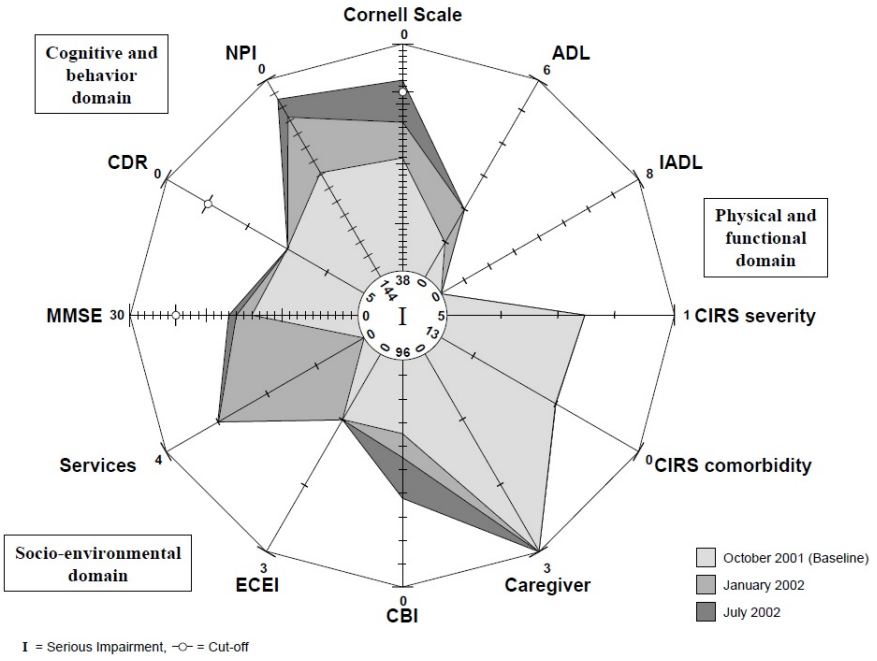
Over the years, several authors have proposed the use of radar charts to organise the visual display of the results of a CGA. This is puzzling because the proposals were made independently of each other, and the authors point out distinct but convergent advantages of the use of this diagram. For example, Vergani *et al.* (2004) proposed the use of a radar chart to show the results of a comprehensive assessment comprising 12 scores, as shown in Figure 3a. Each score belongs to a scale, and each scale is drawn on its own graduated axis. The scales are arranged so that the scores representing favourable health states are plotted closer to the outer border of the chart. Take the ADL scale for instance: it captures the ability to perform tasks of daily living, and ranges from zero to six, with six representing the most favourable state. On the other hand, in the CIRS comorbidity scale, which captures the burden of accumulated chronic diseases and ranges from zero to thirteen, zero represents the most favourable state. The chart also displays cutoffs on some scales, facilitating the identification of domains in which care is needed. According to the authors, the diagram facilitates communication between clinical and administrative staffs, and also stimulates a dialogue that is useful both for clinical and educational reasons. In addition, juxtaposition of results from successive assessments of a patient facilitates the identification of the domains in which a change has occurred, as shown in Figure 3b.

Figure 3 – A radar chart to monitor clinical change over time

(a) Results from an assessment. The scales of each domain are arranged so that scores denoting favourable health states are depicted in the outer border of the chart. Thresholds (cut-offs) plotted on the scales facilitate the identification of health dimensions that need attention.



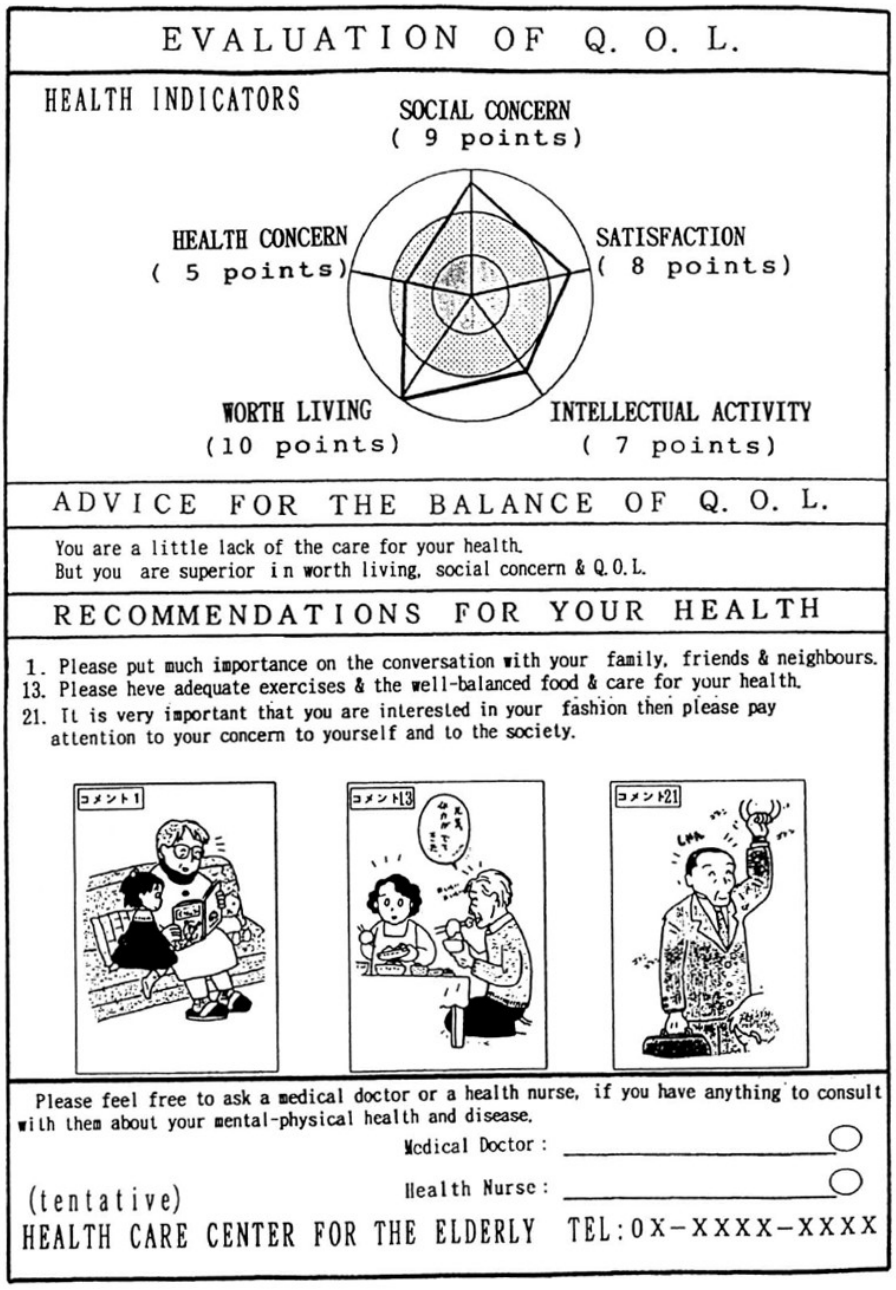
(b) Juxtaposition of three successive assessments of a patient. Changes over time in the monitored health dimensions are easily spotted.



Source: Vergani *et al.* (2004).

Before that work, Shibata *et al.* (1998) employed a radar chart to summarise the assessment of a patient. The questionnaire they used has about 90 health-related items. A scoring procedure converts patient responses into scores on the five domains seen in the radar chart in Figure 4. The chart shows scales without graduation marks, but encodes cutoffs as concentric discs in the middle of the diagram. These facilitate the communication of the health domains in need of attention to the patient. The radar chart is an element of the report produced by an expert system that aimed to promote health-related quality of life of older citizens in Japan. The patient receives a printed copy of this report, which also contains recommendations based on the responses the patient gave to the questionnaire.

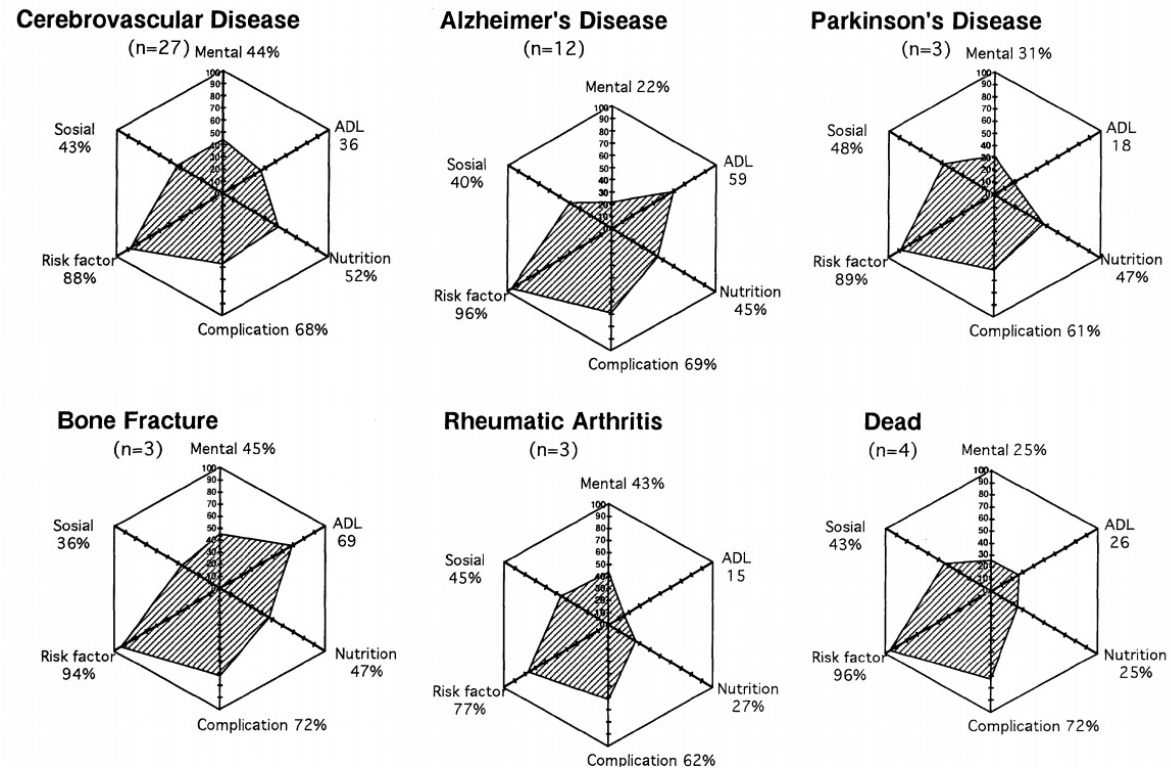
Figure 4 – A radar chart to facilitate communication with patient and family



Source: Shibata *et al.* (1998).

In a different vein, Minemawari and Kato (1999) used radar charts to describe profiles of a set of disabling conditions in older patients admitted to hospital care. As seen in Figure 5, the charts display the scores assigned to groups of patients diagnosed with some relevant condition. A patient assessment comprises six health domains: ADL, mental, social, coronary risk factor, systemic complications, and nutrition. Each domain is represented by a graduated scale, and all scales are normalised to the 0-100 range. One scale per chart has numerical labels for each graduation mark, and each average score is displayed next to its domain label (e.g., "ADL 36" in the top left chart, 36 is the average score for ADL from 27 inpatients). A major difference compared to the use proposed by Vergani *et al.* (2004) is that the diagram seeks to profile a medical condition instead of the health status of an individual. It does so by averaging the assessment scores of a group of patients diagnosed with that condition. The goal of the authors is to foster integrated care by increasing awareness among professionals working in multidisciplinary care teams about how different factors correlate in inpatients with disabling conditions.

Figure 5 – A radar chart to improve communication in the care team

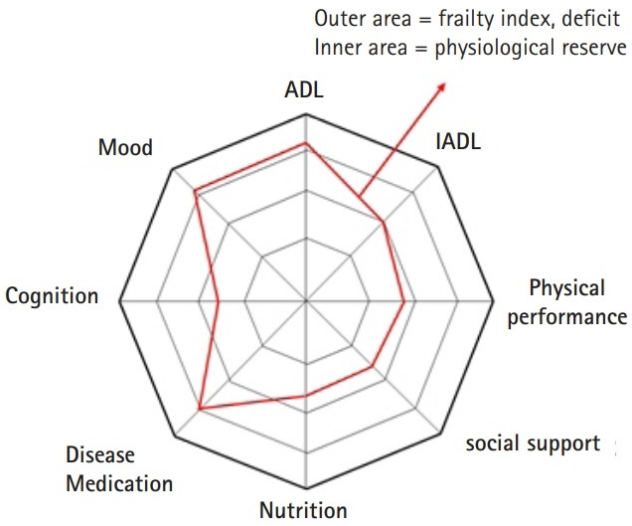


Source: Minemawari and Kato (1999).

Approaching a more conceptual challenge, Jung (2020) has proposed the use of radar charts in the operationalisation of two evasive constructs in the health sciences: physiological reserve and frailty, as seen in Figure 6. Compared to the previous examples, the author innovates in that he explicitly poses a correspondence between these notions and the areas in the diagram. In other words, this is the first clear articulation of an analogy

between area and a multidimensional latent variable that we found in our review. The area within the red polygon (whose vertices correspond to the scores from an assessment) depicts the physiological reserve of an individual, and the area that is external to the polygon and internal to the outer border of the chart (the larger octagon) depicts the notion of frailty being advanced by the author. Similarly to Vergani *et al.* (2004), the scales are graduated (but no numerical labels are displayed), and to Shibata *et al.* (1998), as the author chose not to "fill in" the polygon (which seems counterintuitive in this context). It differs from previous examples in that graduations on one scale are connected with those of neighbouring scales, forming a web-like pattern that visually resembles the discs used by Shibata *et al.* (1998) to represent cutoff levels. According to the author, this diagram "may facilitate communication between healthcare providers to foster shared inter-professional decision-making" and "can make the interpretation of CGA parameters, in order to grasp which domains are impaired, easier, allowing physicians to tackle those with deficits."

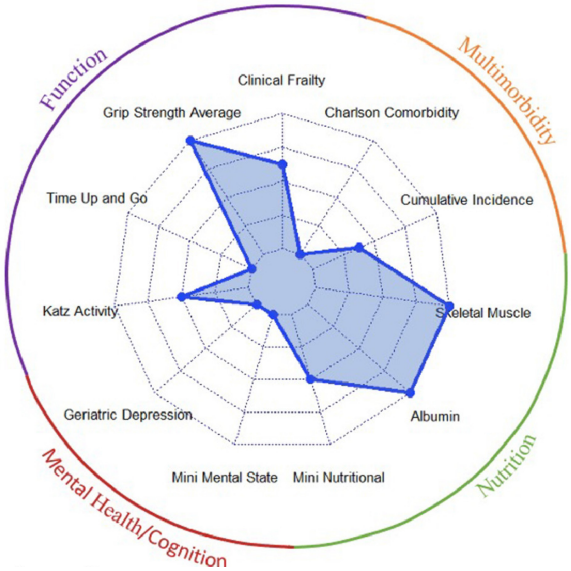
Figure 6 – A radar chart to measure physiological reserve and frailty in older people



Source: Jung (2020).

Finally, Cavanaugh *et al.* (2025) have proposed the use of radar charts to display CGA data in the context of oncological care of older patients. Figure 7 shows a diagram that shares many of the features seen in previous examples. Like in Vergani *et al.* (2004), graduated scales without numerical labels are grouped into domains (as indicated by the outmost circle). Similar to Jung (2020), connections between graduations form a web-like pattern. More importantly, in opposition to Jung (2020), the filled polygon is associated with a notion of frailty instead of physiological reserve, but no meaning is explicitly ascribed to the areas internal or external to the polygon. In this aspect, the diagram is akin to Vergani *et al.* (2004): the polygon represents the "subject's actual recorded condition". According to the authors, the diagram is useful "to guide supportive interventions" and "to facilitate communication of the dominant vulnerability-driving individual risks."

Figure 7 – A radar chart to guide the selection of interventions



Source: Cavanaugh *et al.* (2025).

The five studies we just described were found in a modest but systematic review. Inclusion criteria are primary studies that refer both to CGA and radar charts, and selection was based on titles and abstracts. The Scopus and PubMed databases were searched for such studies and the submitted query follows the pattern: "(`CGA OR elderly questionnaire`) `AND radar chart`", with variables replaced by their corresponding expressions in Table 2. The exclusion criteria were: duplicated studies (multiple proposals by the same author), studies that do not use radar charts to display CGA data, and studies involving populations other than older people. Of the ten studies included in the scope, five were excluded, and the remaining five studies were the focus of our review. The goals and perceived benefits of the use of radar charts to display CGA data are summarised in Table 1. As said before, these are independent proposals (i.e., one work does not cite the others), but the goals and benefits pointed out by the authors are consistent with the goals set by the consensus statement in Brown *et al.* (1988), and also consistent with the goals set by each other.

A final thought: these authors are experts working in a high-stakes environment. According to the classic debate on decision-making theories laid out by Kahneman and Klein (2009), in predictable environments that offer plenty of learning opportunities, professionals often develop intuitive skills. When performing tasks that benefit from such skills, professionals "are often unaware of the cues that guide them." In fact, these cues are believed to trigger pattern recognition processes that, in turn, steer their course of action. In our opinion, the reviewed works were driven by a shared perception of the heuristic value that this visualisation brings to their practice, by giving these cues a diagrammatic expression. This would not be unheard of: logicians, among others, have devised diagrams to materialise and communicate ideas that were once mere abstractions (Legg, 2013).

Table 1 – Review of expected benefits of using radar charts to display CGA data

Goal or perceived benefit	Proposed artefacts per reviewed work				
	Shibata *et al.* (1998)	Minemawari and Kato (1999)	Vergani *et al.* (2004)	Jung (2020)	Cavanaugh *et al.* (2025)
(Section A) Goals of CGA for clinical decision-making					
A1. To improve diagnostic accuracy	expsys	diagram	—	diagram	diagram
A2. To guide the selection of interventions to restore or preserve health	expsys	diagram	diagram	—	**diagram**
A3. To predict health outcomes	—	diagram	—	**diagram**	—
A4. To monitor clinical change over time	—	—	**diagram**	diagram	diagram
(Section B) Perceived benefits of radar chart to display CGA data					
B1. To facilitate communication btw patient and caregiver	**diagram**	—	—	—	diagram
B2. To facilitate communication btw clinical and administrative staff	—	—	diagram	—	—
B3. To stimulate a dialogue that plays both clinical and educational roles	—	**diagram**	diagram	diagram	diagram

Source: The author.

Note: The table has two sections. The four items in Section A come from the statement of the consensus development conference mentioned earlier (Brown *et al.*, 1988). It defined that "the goals of CGA [for clinical decision-making] are: (1) to improve diagnostic accuracy, (2) to guide the selection of interventions ...", and so on. The three items in Section B come from the reviewed studies: the first one from Shibata *et al.* (1998) and the other two from Vergani *et al.* (2004). These correspond to perceived benefits that the use of radar charts brings to the clinical practice. The remaining columns refer the five works we reviewed, in which the authors proposed the use of radar charts to display CGA data. The term "diagram" in a cell indicates that the authors (in the column) find that the diagram they proposed serves the goal or creates the benefit described in that row. The term "expsys" stands for expert system, a precursor to modern recommender systems. A term is print in bold to indicate that the authors highlighted the corresponding goal or benefit in the paper's abstract, which we interpret as an expression of importance given by the authors.

Table 2 – Variables used in searching for CGA studies using radar charts

Variable	Expression
`CGA`	"comprehensive geriatric assessment" `OR` "comprehensive geriatric evaluation" `OR` "comprehensive gerontological assessment" `OR` "comprehensive gerontological evaluation" `OR` "multidimensional geriatric assessment" `OR` "multidimensional geriatric evaluation" `OR` "multidimensional gerontological assessment" `OR` "multidimensional gerontological evaluation" `OR` "multidisciplinary geriatric assessment" `OR` "multidisciplinary geriatric evaluation" `OR` "multidisciplinary gerontological assessment" `OR` "multidisciplinary gerontological evaluation"
`elderly questionnaire`	"questionnaire" `AND` ("health" `OR` "healthcare" `OR` "geriatrics" `OR` "gerontological" `OR` "quality of life" `OR` "well-being" `OR` "intrinsic capacity" `OR` "functional ability" `OR` "daily living") `AND` ("elder" `OR` "elderly" `OR` "senior" `OR` "older person" `OR` "older adult" `OR` "older patient" `OR` "older people" `OR` "older population")
`radar chart`	"radar chart" `OR` "radar diagram" `OR` "radar graph" `OR` "radar plot" `OR` "web chart" `OR` "web diagram" `OR` "web graph" `OR` "web plot" `OR` "coweb chart" `OR` "coweb diagram" `OR` "coweb graph" `OR` "coweb plot" `OR` "spider chart" `OR` "spider diagram" `OR` "spider graph" `OR` "spider plot" `OR` "star chart" `OR` "star diagram" `OR` "star graph" `OR` "star plot" `OR` "polar chart" `OR` "polar diagram" `OR` "polar graph" `OR` "polar plot" `OR` "Kiviat chart" `OR` "Kiviat diagram" `OR` "Kiviat graph" `OR` "Kiviat plot"

Source: The author.

2.3 The measurement of health, psychometrics, and factor analysis

Measurement is seen by many as a hallmark of modern science, but "there is little consensus among philosophers as to how to define measurement, what sorts of things are measurable, or which conditions make measurement possible", argues Tal (2020). This statement reflects unsettled disputes about measurement in the social sciences and in the part of health sciences that overlaps with psychology. In these branches, researchers are often faced with the challenge of measuring variables that cannot be directly observed, such as life satisfaction or functional independence (Philippi, 2023). The difficulties emerge early in the process of developing a systematic measurement method, in the conceptualisation stage, because even everyday concepts may turn out to be notoriously elusive to define. For example, most people will take only a couple of seconds to answer "How satisfied are you with your health?" but take the definition of health that appears in the preamble to the 1947 Constitution of the WHO: "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (Hausman, 2015). How many people would weigh all these aspects in their answer? Who should and when?

This definition of health, while aspirational, presents a huge challenge to researchers: How should one measure something as intangible as a "state of complete [...] well-being?" In the past, social researchers looked to the measurement methods of the natural sciences for principles that could improve their measurement practices (Michell, 1999). Nowadays, many are sceptical about the efficacy of this research programme, and for historical reasons, the gerontological research community has predominantly subscribed to psychometric methods to meet the challenge of measuring the human attributes that are substantive to their theories, and to the methods of factor analysis in particular[2] (Shenk *et al.*, 2001).

Borsboom and Molenaar (2015) define psychometrics as "a scientific discipline concerned with the question of how psychological constructs (e.g., intelligence, neuroticism, or depression) can be optimally related to observables (e.g., outcomes of psychological tests, genetic profiles, neuroscientific information)." In other words, the central question of psychometrics, which is of direct relevance to gerontological research, is how to connect theory to observations. In the factor-analytic approach, this connection can be formalised by a class of statistical models known as structural equation models (Bollen, 1989).

[2] The Gerontological Society of America (GSA), founded in 1945, is one of the oldest associations of gerontologists worldwide. Back then, psychometric methods based on factor analysis were reasonably mature and widely used in the social sciences (Bollen, 1989). The major alternative models in psychometrics, the IRT models, would only gain momentum with the Rasch model in the 1960s, and clinimetrics, the major alternative tradition, would only emerge in the 1980s. In a sense, factor analysis was the only game in town, and one could argue that it is the predominant choice even today in human sciences research. For example, the GSA publishes The Gerontologist journal since 1961. A search for the term "factor analysis" in this journal returns 159 articles, 17 for "IRT", and one for "clinimetric" (as of October, 2025). Similar numbers are reported for works in psychology & education by Holt (2014, Ch. 2).

An interesting characteristic of these models is that the relationships among latent variables are specified separately from the relationships between each latent variable and its indicators, as illustrated in Figure 8a. The latent variables represent the theoretical constructs being investigated, which together with their relationships form the latent variable (sub)model. A measurement (sub)model comprises a latent variable, its indicators, and their relationships. Indicators are observed variables that are related to the data that the researcher collects to assess the plausibility of the relationships in the model.

In a nutshell, once a structural equation model is fitted, the researcher evaluates the extent to which the model provides an adequate description of the covariances observed in the data. Guided by normative fit indices, the researcher determines if the model adequately reproduces the observed covariances (and thus supports the specified relationships), or if an alternative model should be considered (Holt, 2014). When a model achieves an adequate fit and its parameters agree with the researchers' expectations, its measurement models can potentially be reused in other settings to measure the same latent variable.

There is a vast literature on structural equation models that covers how models should be specified, which algorithms can be used to estimate the parameters of a model, and how the researcher should evaluate and report obtained results based on generally accepted normative indices[3]. However, we focus on measurement models because of their importance as building blocks in the development of psychometric scales. This means that this structure is a common element of several psychometric instruments, a fact that we later exploit to propose a recommendation model for data collected with such instruments.

To reduce ambiguity, we clarify our usage of some technical terms in the following. In the measurement model shown in Figure 8b, η represents a latent variable. In the literature, there are several definitions of what a latent variable is, but we use the term in the following sense: a latent variable is a random variable that represents a psychological attribute; this attribute is assumed to exist independently of measurement, but (presently) cannot be measured directly without substantial measurement error (Bollen; Pearl, 2013; Holt, 2014). Although a latent variable is hardly accessible to direct measurement, the fact that it acts as a common determinant of a set of variables that are more amenable to direct

[3] In the literature on recommender systems, Knijnenburg and Willemsen (2015) described how structural equation models can be used to design and analyse the outcome of user experiments. We refer the reader interested in a thorough mathematical approach to structural equation models to the classic text of Bollen (1989). A concise presentation of factor-analytic models with a survey on how they have been employed by researchers has been written by Holt (2014). For those interested in a critical discussion about the philosophical aspects of psychometrics, we recommend the book by Borsboom (2005), which is supplemented by his video lecture (here) commissioned by the Psychometric Society. A somewhat opposite view has been presented by Michell (1999), in which the author retraces many historical developments around measurement practices leading to the birth of psychometrics as an academic discipline int the 1930s. For the Portuguese-speaking reader, an introductory text on factor analysis with worked out examples has been published by Aranha and Zambaldi (2008).

measurement allows us some level of access to it (Borsboom, 2005). The arrows that depart from η towards the indicators $x_1 \dots x_3$ represent its action as a common determinant. Note that $x_1 \dots x_3$ are represented as latent variables in Figure 8a and as observed variables in Figure 8b. This reflects a subtle duality: once we measure the indicators of a latent variable and gain some access to it, the variable ceases to be "hidden" and can be treated as a manifest variable. This latter term has multiple synonyms in the literature (e.g., indicator, observed variable, observable variable), and we use them interchangeably with this meaning: a manifest variable is a random variable that is more amenable to direct measurement compared to the latent variables in a structural model.

An example to embody these ideas: suppose that x_1 in Figure 8a is a latent variable that represents psychological well-being. Then the variable q_1 could represent the response to an item in a questionnaire asking the respondent to rate the statement "How well are you able to concentrate?" on a five-point Likert scale ranging from "Not at all" to "Extremely". Once the respondent answers the items related to $q_1 \dots q_4$ (and assuming that the parameters $\gamma_{11} \dots \gamma_{14}$ are known), the collected responses can be used to compute a score for x_1 (McNeish; Wolf, 2020). This causes an upstream effect: x_1 can now be used as an indicator of η, as seen in Figure 8b. In a sense, the measurement model of the latent variable x_1, which comprises x_1, $q_1 \dots q_4$, $\gamma_{11} \dots \gamma_{14}$, and $\delta_1 \dots \delta_4$, works as a scale for x_1 in much the same way as the five-point Likert scale ranging from "Not at all" to "Extremely" is a scale for q_1. In this usage, the term scale means "a system for ordering test responses in a progressive series, so as to measure a trait, ability, attitude, or the like", which corresponds to a definition found in the APA's online dictionary. In the same vein, we use the term instrument to refer to a system by which researchers assess or gather data about study participants. For example, suppose that η in Figure 8a is a latent variable that represents a health-related notion of quality of life. Then the questionnaire containing the items related to $q_1 \dots q_{12}$, together with a scoring procedure, is an instrument[4], and we give a special name to the latent variables $x_1 \dots x_3$: they are the domains of the instrument.

Now that we have set up the appropriate nomenclature, let's have a look at how structural models are estimated. From the path diagram in Figure 8b, one can deduce that the score that the i-th subject obtains for the k-th domain, namely x_{ik}, depends on the her position on the measurand (η_i), the effect the measurand exerts on its indicators (λ_k), and a measurement error (ϵ_{ik}). This is formalised by a structural equation (Holt, 2014):

$$x_{ik} = \lambda_k \eta_i + \epsilon_{ik}. \tag{2.1}$$

In fact, the whole diagram is a pictorial representation of the system of structural equations

[4] In the literature, the term inventory is sometimes used to denote something similar to an instrument, but we make a distinction: the structural model used to develop an instrument has a top variable, as illustrated in Figure 8a. We use the term inventory to denote a questionnaire that has validated scales, but lacks evidence that they measure a common cause.

shown to the right of the diagram. Suppose that a researcher is creating a new instrument that surveys d domains, each with a mature scale already available. The researcher gathers data from m subjects, computes their scores for each domain, and organises the scores into a (d, m)-matrix $\mathring{X}$. The structural equations indicate that the matrix $\mathring{X}$, once centred, could be rewritten as $X = \Lambda H' + \varepsilon$, with H (uppercase η) as a $(m, 1)$-vector with the scores of the m subjects on the latent variable η, Λ as a $(d, 1)$-vector with the effect parameters, and ε as an (d, m)-matrix with measurement errors[5]. From this matrix equation, one can deduce the (d, d)-covariance matrix of X, namely $\Sigma = \mathbb{E}[XX'] = \Lambda\phi\Lambda' + \Theta$, with ϕ as the scalar variance of the latent variable η, and Θ as a (d, d)-matrix with error covariances. The estimation process seeks for the values of the model parameters (Λ, ϕ, Θ) that minimise the "distance" between the predicted covariance matrix Σ and the sample covariance matrix calculated from $\mathring{X}$, subject to a number of constraints, such as $\mathbb{E}[H] = 0$, $\mathbb{E}[\varepsilon_k] = 0$, $\mathrm{Cov}(H, \varepsilon_k) = 0$ for all k, and $\mathrm{Cov}(\varepsilon_k, \varepsilon_l) = 0$ for $k \neq l$, among possibly others[6].

An interesting result of this dynamics, which we later exploit in the data preparation procedure of our recommendation model, is that if a model is set so that $\lambda_k > 0$ for all k (as is usual) and the model fits the data well, then the sample covariance matrix should have only positive entries (Bollen, 1989, p.22):

$$\begin{aligned}
\mathrm{Cov}(\mathring{X}_k, \mathring{X}_l) &= \mathbb{E}[\,(\mathring{X}_k - \mathbb{E}[\mathring{X}_k]\,)\,(\mathring{X}_l - \mathbb{E}[\mathring{X}_l]\,)'\,] = \mathbb{E}[X_k\, X_l'] \\
&= \mathbb{E}[(\lambda_k H' + \varepsilon_k)(\lambda_l H' + \varepsilon_l)'] = \mathbb{E}[(\lambda_k H' + \varepsilon_k)\,(\lambda_l H + \varepsilon_l')] \\
&= \mathbb{E}[\lambda_k\lambda_l H'H + \lambda_k H'\varepsilon_l' + \lambda_l\varepsilon_k H + \varepsilon_k\,\varepsilon_l'] = \lambda_k\lambda_l\phi > 0. \qquad (2.2)
\end{aligned}$$

Once the model parameters are estimated, a number of statistical indices of fit can be computed. The literature suggests acceptable ranges for several of these indices, and researchers rely on these ranges to assess the quality of an instrument based on found (or reported) results. Of particular interest to our project is the instrument's reliability. According to Borsboom and Molenaar (2015), "a measurement instrument is reliable to the extent that it yields the same outcomes when applied to persons with the same standing of the measured attribute under the same circumstances," and Hayes and Coutts (2020) remind us that reliability is "a property of the data generated by the instrument when applied to a specific population rather than a property of the scale itself." In general, reliability is conceptualised as a signal-to-noise ratio that quantifies the amount of random measurement error that exists in a set of measurements. For congeneric models such as the one in Figure 8b, the McDonald's ω is a recommended index (McNeish, 2018):

$$\omega = \frac{\left(\sum_k \lambda_k\right)^2}{\left(\sum_k \lambda_k\right)^2 + \sum_k \mathrm{Var}(\varepsilon_k)}, \qquad (2.3)$$

[5] In this section, we depart from our notation, in which $\mathring{X}$ is an (m, d)-matrix, to use that of the structural modelling literature. Note, however, that $\mathring{X}_k$ still refers to the k-th row of $\mathring{X}$.

[6] For more details, consult the texts by Holt (2014, p.10) and Bollen (1989, p.35).

where ε_k is the k-th row of the measurement errors matrix; assuming that $\text{Var}(\varepsilon_k)$ is finite for all k, it must be the case that $\omega \in (0, 1]$. To paraphrase Hayes and Coutts (2020), the higher the reliability, the more we can trust that the differences between people in their scores on the latent variable η are an accurate reflection of actual individual differences in the attribute represented by η. Ideally, researchers should be adamant about using instruments with demonstrated reliability in the population being investigated.

If an instrument demonstrates that it has the desirable psychometric properties to support its proposed uses, and one of these uses is establishing baseline scores in clinical trials (i.e., sensitive to within-subject difference), then the instrument's scoring procedure gains a prominent role. For the congeneric model in Figure 8b, there are two scoring procedures that are predominantly used: the sum-score and the optimally weighted score (McNeish; Wolf, 2020). The first is the sum of the domain scores obtained by an individual:

$$\hat{\eta}_i = \sum_k \mathring{x}_{ik}. \tag{2.4}$$

Some researchers argue that this is not the best scoring method for congeneric models because it assumes that all indicators reflect the latent variable being measured equally (i.e., it assumes that $\lambda_k = C$ for all k in Figure 8b, with $C \in \mathbb{R}$). An alternative is provided by the optimally weighted score because it takes into account the information about the effects that the latent variable exerts on the indicators (Thissen *et al.*, 1983, Eq. 9 and 10):

$$\hat{\eta}_i := \frac{\sum_k \frac{\lambda_k}{\sigma_k^2}(\mathring{x}_{ik} - \tilde{x}_k)}{1 + \sum_k \frac{\lambda_k^2}{\sigma_k^2}} \propto \sum_k \frac{\lambda_k}{\sigma_k^2}(\mathring{x}_{ik} - \tilde{x}_k) \overset{\propto}{\approx} \sum_k \frac{\lambda_k}{\sigma_k^2}\mathring{x}_{ik} =: \hat{\eta}_i^*, \text{ with} \tag{2.5}$$

$$(a \propto b) \implies \exists\, C \in \mathbb{R}, C > 0 : a = Cb,$$

$$(a \overset{\propto}{\approx} b) \implies \exists\, C_0, C_1 \in \mathbb{R}, C_1 > 0 : a = C_1 b + C_0,$$

$$\tilde{x}_k := \mathbb{E}[\mathring{X}_k], \text{ and } \sigma_k^2 := \text{Var}(\varepsilon_k).$$

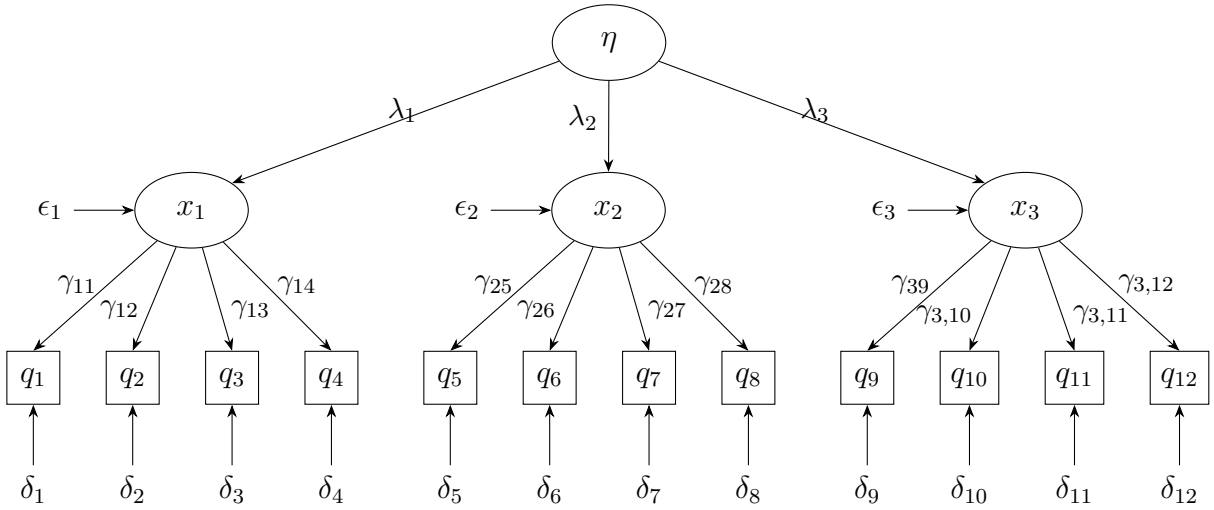
In summary, this means to say that $\hat{\eta}_a > \hat{\eta}_b \iff \hat{\eta}_a^* > \hat{\eta}_b^*$. In addition, if we assume that $\sigma_k^2 = C \in \mathbb{R}$ for all k, then Equation 2.5 simplifies to a weighted sum of the item responses:

$$\hat{\eta}_i^* := \sum_k \frac{\lambda_k}{C}\mathring{x}_{ik} \propto \sum_k \lambda_k \mathring{x}_{ik} =: \hat{\eta}_i^\dagger, \tag{2.6}$$

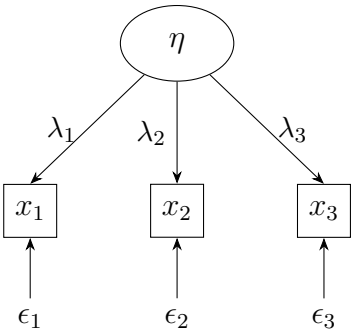
Note that the sum-score in Equation 2.4 is a special case of Equation 2.6 with the effect coefficients set to one. Both scoring methods have limitations, and the relative advantages of one over another are a subject of debate even today (McNeish; Wolf, 2020; Widaman; Revelle, 2023; McNeish, 2023; Widaman; Revelle, 2024; Sijtsma; Ellis; Borsboom, 2024a; McNeish, 2024; Sijtsma; Ellis; Borsboom, 2024b). The three instruments that we will cover in the next section use a scoring procedure that is equivalent to the sum-score in Equation 2.4, but we will later exploit the result in Equation 2.6 to establish a relationship between scores on the latent variable η and the geometry of the visualisations seen in Section 2.2.

Figure 8 – A structural equation model represented by a path diagram

(a) The path diagram of a structural equation model with latent variables



(b) A congeneric measurement model for the top latent variable



Structural Equations

$$\begin{cases} x_{i1} = \lambda_1 \eta_i + \epsilon_{i1} \\ x_{i2} = \lambda_2 \eta_i + \epsilon_{i2} \\ x_{i3} = \lambda_3 \eta_i + \epsilon_{i3} \end{cases}$$

Covariance Matrix

$$X := \Lambda H' + \varepsilon \implies X' = H\Lambda' + \varepsilon'$$

$$XX' = \Lambda H' H \Lambda' + \Lambda H' \varepsilon' + \varepsilon H \Lambda' + \varepsilon\varepsilon'$$

$$\Sigma = \mathbb{E}[XX'] = \Lambda\phi\Lambda' + \Theta$$

Source: The author.

Legend: A symbol enclosed by an ellipse represents a latent variable (e.g., η); a symbol enclosed by a square represents an indicator (or manifest variable: the q_j variables); an unenclosed symbol represents an error variable (all δ_j and ϵ_j variables, which are unobserved); a straight arrow from A to B indicates causation and should be read as "A causes B" (Bollen, 1989). A more nuanced interpretation of the straight arrow is given by Borsboom (2005): "variation in the latent variable precedes variation in the indicators". Other visual types commonly used in path diagrams are ignored in our discussion.

Notes: Figure 8a shows an adaptation of the path diagram for the second order confirmatory factor analysis (CFA) model from Bollen (1989, p. 315). This model has been employed in prominent models used in gerontological research, including some instruments described in the next section. In this diagram, the latent variable model is composed of the subgraph induced by the latent variables, namely $\eta, x_1 \dots x_3$. It encodes the fact that the second order variable η causes variations in the first order variables $x_1 \dots x_3$, but the latter do not influence each other. There are three measurement models, one for each of the first order variables (i.e., x_1 to x_3). A measurement model is composed of the subgraph induced by a given latent variable and its corresponding indicators. Figure 8b focus on the collapsed measurement model for the latent variable η to show that its path diagram is a pictorial representation of a system of structural equations, and that the hypothetical covariance matrix Σ is derived from this system.

2.4 Three psychometric instruments used in gerontological research

In recent years, there has been a shift in focus in health care: patients and their families want to be involved in the decision-making process. A central piece of this patient-centred care movement, which enables care professionals to gain insight about the patients' perspective about their health, is the development of reliable patient reported outcome measures (PROMs) and their incorporation into the traditional ways of measuring health and the effects of treatment on the patient (Krabbe, 2016). These measures also play a significant role in initiatives to promote early interventions as a more effective means of optimising the health trajectories of individuals from a longitudinal perspective. The techniques reviewed in the previous section are instrumental to the success of these efforts.

In this section, the structure of three such instruments is briefly described: the WHOQOL-BREF instrument for quality of life, the WHO's instrument for intrinsic capacity, and the AMPI-AB instrument for fragility. For each instrument, we briefly review the major concerns that lead to its development, the factor-analytic structure of its underlying model, and its scoring procedure. These instruments will be used in later chapters to describe the datasets used in the evaluation of our proposed model.

2.4.1 Measuring quality of life

The WHOQOL-BREF instrument is a shorter version of the WHOQOL-100, which was developed by the WHO in the 1990s in response to an increasing demand for subjective measures of health and well-being that could be applied to populations worldwide (including Brazil). The WHO defines quality of life as "an individual's perception of their position in life in the context of the culture and value systems in which they live, and in relation to their goals, expectations, standards and concerns" (Skevington; Lotfy; O'Connell, 2004).

The instrument has a questionnaire with 26 items that survey four domains: physical health (7 items), psychological well-being (6 items), social relationships (3 items), and environmental quality (8 items). The remaining two items survey the general perception of health and quality of life (Orley *et al.*, 1996). The respondent rates their agreement with the statement of an item on a five-point Likert scale, whose labels are adapted for each domain, as shown in Table 3. Responses are encoded as numeric values from one to five.

The instrument was developed using a second order confirmatory factor analysis model similar to the one illustrated in Figure 8a. The 2012 revision of the WHOQOL manual reports the following effect coefficients: 0.84 for the physical domain, 1.0 for the psychological domain, 0.85 to the social domain, and 0.77 for the environment domain (World Health Organization, 2012). The reliability was calculated for each domain separately using the Cronbach's alpha (instead of the McDonald's omega) and found to demonstrate good internal consistency (values varying between 0.66 and 0.84).

Table 3 – Questionnaire of the WHOQOL-BREF instrument

Ord	Domain	Item phrasing	Scale (extremes)
1	general	How would you rate your quality of life?	Very poor to Very good
2	general	How satisfied are you with your health?	Very dissatisfied to Very satisfied
3*	physical	To what extent do you feel that physical pain prevents you from doing what you need to do?	Not at all to An extreme amount
4*	physical	How much do you need any medical treatment to function in your daily life?	Not at all to An extreme amount
5	psychological	How much do you enjoy life?	Not at all to An extreme amount
6	psychological	To what extent do you feel your life to be meaningful?	Not at all to An extreme amount
7	psychological	How well are you able to concentrate?	Not at all to Extremely
8	environment	How safe do you feel in your daily life?	Not at all to Extremely
9	environment	How healthy is your physical environment?	Not at all to Extremely
10	physical	Do you have enough energy for everyday life?	Not at all to Completely
11	psychological	Are you able to accept your bodily appearance?	Not at all to Completely
12	environment	Have you enough money to meet your needs?	Not at all to Completely
13	environment	How available to you is the information that you need in your day-to-day life?	Not at all to Completely
14	environment	To what extent do you have the opportunity for leisure activities?	Not at all to Completely
15	physical	How well are you able to get around?	Very poor to Very good
16	physical	How satisfied are you with your sleep?	Very dissatisfied to Very satisfied
17	physical	How satisfied are you with your ability to perform your daily living activities?	Very dissatisfied to Very satisfied
18	physical	How satisfied are you with your capacity for work?	Very dissatisfied to Very satisfied
19	psychological	How satisfied are you with yourself?	Very dissatisfied to Very satisfied
20	social	How satisfied are you with your personal relationships?	Very dissatisfied to Very satisfied
21	social	How satisfied are you with your sex life?	Very dissatisfied to Very satisfied
22	social	How satisfied are you with the support you get from your friends?	Very dissatisfied to Very satisfied
23	environment	How satisfied are you with the conditions of your living place?	Very dissatisfied to Very satisfied
24	environment	How satisfied are you with your access to health services?	Very dissatisfied to Very satisfied
25	environment	How satisfied are you with your transport?	Very dissatisfied to Very satisfied
26*	psychological	How often do you have negative feelings such as blue mood, despair, anxiety, depression?	Never to Always

Source: The author.

Note: The first column indicates the order in which the item appears in the questionnaire. An asterisk after the number of the item means that it is negatively worded, so it must be reversed during the scoring procedure. Note that items that survey distinct domains are interwoven throughout the questionnaire. In fact, they are organized into eight sequential blocks of items that share the same rating scale.

The instrument has a formal scoring procedure for the domains but does not specify a single score for quality of life. This is indicative that the four domain scores should be interpreted as a (multidimensional) profile of quality of life instead of a unidimensional measure. The scoring rules are: (a) reverse the score of the three negatively phrased items, and (b) average the responses obtained for each domain and multiply the average by four (to produce scores that are compatible with the WHOQOL-100 instrument). The averaged score is proportional to the sum-score from Equation 2.4 and therefore both methods are equivalent for the purpose of ordering of people on any given score. The instrument also has rules to handle missing data, but it suffices to say that that scoring procedure works even if one response per domain is missing.

2.4.2 Measuring intrinsic capacity

As we discussed briefly in Section 2.1, the WHO has been advancing a change in national public health policies in response to the fast demographic changes that are underway worldwide. This change consists in shifting the focus on disease management to an approach that seeks to preserve health of individuals to the older age. According to Carvalho *et al.* (2017), "instead of trying to manage an array of diseases and treat specific symptoms in a disjointed fashion, interventions should be prioritised in ways that optimise trajectories of older people's physical and mental capacities." An essential component of this strategy is the development of an instrument to assess intrinsic capacity, so that public policies can be evaluated for their effectiveness in preserving intrinsic capacity of populations served by national health systems across the life course of their members.

In this framework, intrinsic capacity is defined as "the composite of all the physical and mental capacities of an individual", which is essential for their everyday functioning (World Health Organization, 2015). This construct has been operationalised by the ICOPE initiative mentioned earlier (Tavassoli *et al.*, 2022; Hsiao; Chen, 2024), and encompasses five domains: sensory function, cognition, vitality, locomotion, and psychological well-being.

A recent study to assess the psychometric properties of an instrument to assess intrinsic capacity of Brazilian populations was conducted by Aliberti *et al.* (2022). The authors used data from the baseline assessment of the ongoing ELSI-Brazil project to compute intrinsic capacity profiles for the surveyed populations. The resulting questionnaire, which is shown in Table 4, is composed of items that were submitted to the study participant (e.g., Who is the current president of Brazil?) and physical tests (e.g., time the participant takes to walk three metres at the usual pace). The first column of the table indicates the code(s) of the item(s) in the original ELSI's questionnaire (Lima-Costa *et al.*, 2018). An asterisk after the code indicates it must be reversed during the scoring procedure. Note that several items in this questionnaire come from scales widely used in healthcare research, e.g., the 8-items CES-D instrument for depression (items r2 to r9), and the physical tests.

Table 4 – Questionnaire to assess the domains of intrinsic capacity

Src	Domain	Item phrasing or Test	Scale
q7:10	cognitive	Orientation for year, month, day, and day of the week.	count of correct responses
q13	cognitive	Delayed recall of 10 common words.	count of correct responses
q18	cognitive	What do people usually use to cut a paper?	count of correct responses
q19	cognitive	What is the plant that has a long and green leaf that gives a yellow and long fruit and that we peel to eat it?	count of correct responses
q20	cognitive	Who is the current president of Brazil?	count of correct responses
q21	cognitive	Who is the current vice-president of Brazil?	count of correct responses
q14	cognitive	Distinct animals in 60 seconds.	count of correct responses
r2*	psychological	During the last week, have you felt depressed most of the time?	Yes, No
r3*	psychological	During the last week, have you felt like things were harder?	Yes, No
r4*	psychological	During the last week, have you felt like your sleep wasn't restful most of the time?	Yes, No
r5	psychological	Did you feel happy most of the time?	Yes, No
r6*	psychological	Did you feel lonely most of the time?	Yes, No
r7	psychological	Did you enjoy or enjoy life most of the time?	Yes, No
r8*	psychological	Did you feel sad most of the time?	yes, no
r9*	psychological	Did you feel like you couldn't get things done?	Yes, No
n74*	psychological	How would you evaluate the quality of your sleep?	Excellent to Very poor
n75*	psychological	During the last month, have you taken any sleeping pill?	No, Less than once a week, Btw 1-2 times a week, 3+ times a week.
n16*	sensory	How do you evaluate your hearing (even when using a hearing device)?	Excellent to Very poor
n6*	sensory	How good is your eyesight (even when using glasses or contact lenses) for seeing things at a distance, like recognizing a friend across the street?	Excellent to Very poor
n7*	sensory	How good is your eyesight (even when using glasses or contact lenses) for seeing things up close like reading ordinary newspaper print?	Excellent to Very poor
mf33-mf38*	locomotor	Time to walk three meters at the usual pace with or without assistive devices.	in meter/sec, discretised
mf30-mf32	locomotor	Balance test from the Short Physical Performance Battery.	in sec, adjusted per age group, discretised
mf27-mf29	vitality	Handgrip strength of the dominant hand evaluated using a Sachan handheld dynamometer.	in kg, adjusted per sex and body mass, discretised
n69* n70*	vitality	Unintentional weight loss 3kg or more during the last 3 months.	Yes, No
n72*	vitality	In the past week, how often did you feel that you could not carry things forward?	Never or rarely (less than 1 day), Very few times (1-2 days), Sometimes (3-4 days), Most of the time
n73*	vitality	In the past week, how often did the routine activities require a major effort to be completed?	Never or rarely (less than 1 day), Very few times (1-2 days), Sometimes (3-4 days), Most of the time

Source: The author.

An early version of this instrument was conceived using a second order confirmatory factor analysis model similar to the one illustrated in Figure 8a. Carvalho *et al.* (2017) reported the following coefficients: 0.45 for sensory function, 0.64 for cognition, 0.59 for vitality, 0.95 for locomotion, and 0.57 for psychological well-being. No assessments of reliability were reported, as is usual in the initial stages of development of a new instrument. Following a more recent trend, Aliberti *et al.* (2022) used a bifactor model to evaluate the instrument. They concluded that the model revealed satisfactory robust goodness-of-fit indices compared to other factor models. No standard scoring procedure was found in the surveyed articles, as would be expected as the instrument is still under development. It must be noted, however, that the wide variety of scales must be taken into account in the scoring procedure, to ensure that indicators of a given domain do not dominate the others because the numeric values (arbitrarily) assigned to its levels vary over different ranges.

2.4.3 Measuring frailty

In response to an increasingly older population in Brazil, and in agreement with the need for improvements in national health care systems to provide integrated care for this population, as advanced by the WHO, the Municipal Health Department of the city of São Paulo has instituted the Elderly Caregiver Program (PAI) programme. This is a home care programme that offers clinically frail and socially vulnerable older people "the services of health care professionals and professional caregivers aiming at rehabilitation, maintenance/improvement of selfcare, and socialization" (Andrade *et al.*, 2020).

The Multidimensional Evaluation of Older People in Primary Care (AMPI-AB) instrument is used by the PAI programme as one of the criteria to include people to be assisted. This instrument is under development, but a recent work has shown that most of its items cluster around five domains: cognitive (3 items), activities of daily living (ADL) (4 items), instrumental activities of daily living (IADL) (4 items), oral health (4 items), and morbidities (3 items) (Andrade, 2019, Table 5). The respondent rates their agreement with the statement of an item using several scales, as shown in Table 5. To the best of my knowledge, a confirmatory factor analysis was not performed on this instrument, but Andrade (2019) reports the results of an exploratory factor analysis in which the target construct seems to be some notion of frailty, since the score an individual obtains is used to classify her into one of three classes: healthy (0 to 5 points), pre-frail (6 to 10 points), or frail (more than 10 points). The instrument uses the sum-score from Equation 2.4. With regard to reliability, Andrade (2019) reports a value of 0.79 for the Cronbach's α and of $0,78$ for the McDonald's ω, which indicates an acceptable level for early stages of research. However, several measures of closeness to unidimensionality suggest that the data collected with this instrument will fit poorly to a congeneric model: a value of 0.84 is reported for UNICO (Unidimensional Congruence), among others.

Table 5 – Questionnaire of the AMPI-AB instrument

Ord	Domain	Item phrasing	Scale (extremes)
2*	morbidities	In general, compared to other people your age, would you say your health is:	Very poor to Very good
4	morbidities	Have you had/do you have any of the conditions below?	count of reported items from a predefined list
5	morbidities	How many medications do you take daily?	count of reported items
11a	cognitive	Has a family member or friend told you that you are becoming forgetful?	Yes, No
11b	cognitive	Has your forgetfulness worsened in recent months?	Yes, No
11c	cognitive	Is your forgetfulness preventing you from performing any daily activities?	Yes, No
13a	ADL	Do you need help getting out of bed?	Yes, No
13b	ADL	Do you need help getting dressed?	Yes, No
13c	ADL	Do you need help eating?	Yes, No
13d	ADL	Do you need help bathing?	Yes, No
10c*	IADL	Can you walk 400 meters (approximately four blocks)?	Yes, No
10d*	IADL	Can you sit or stand up without difficulty?	Yes, No
14a	IADL	Do you need help performing activities outside the home?	Yes, No
14b	IADL	Do you need help managing money (paying bills, checking change, going to the bank, etc.)?	Yes, No
17a	oral health	If you wear dentures, are they poorly fitted?	Yes, No
17b	oral health	Do you have trouble chewing?	Yes, No
17c	oral health	Do you have trouble swallowing?	Yes, No
17d	oral health	Have you stopped eating any foods because of problems with your teeth or dentures?	Yes, No

Source: The author.

Note: The first column indicates the order in which the item appears in the questionnaire. An asterisk after the number of the item means that it is negatively worded, so it must be reversed during the scoring procedure.

2.5 Summary and closing remarks

This chapter surveyed how gerontologists depend on CGA in primary care, a tool that evaluates multiple biopsychosocial domains known to correlate with functional ability in older populations. We briefly described how psychometrics provides the methodological framework to build such tools and detailed some real-world examples of psychometric instruments used for clinical research and practice in gerontology. These results, especially Equations 2.2 and 2.6, will be used later to describe the expected structure of psychometric data, in the context of training and evaluating machine learning models. The chapter also reviewed recurring calls in the gerontological literature to use radar charts to display the results of a CGA. We conducted a systematic review and then analysed and compared the five studies we found. This analysis will later be used to show a relationship between the geometry of a radar chart and the congeneric model in Figure 8b based on Equation 2.6.

3 RELATED WORK

In this chapter, we review the literature on healthcare recommender systems that target older people. The review aims to clarify the gap between current system designs and applications and the minimal requirements they should meet to function as tools to assist gerontologists create personalised care plans in primary care. In our view, these requirements must include the use of CGA data to describe patients, as well as the provision of faithful explanations directed at the attending care professional.

Two other important topics are covered in this chapter. The literature on multilabel classification and label ranking tasks is reviewed because these tasks can be used to formalise the task a gerontologist performs when creating a personalised care plan. We review the nomenclature, the main technical approaches to solving the tasks, and the benchmark datasets with their standard evaluation methodologies. Finally, we also briefly review the literature on visualisation research to summarise current results on the types of visualisation that best support the completion of decision-making tasks. These results will be used later to inform our discussion about the measurement of interpretability.

3.1 Applications of recommender systems in healthcare

Recommender systems have demonstrated their usefulness by their widespread application in a variety of domains: from consumer goods to travel and entertainment services, from news to academic articles and books, from friends and dates to job offers (Ricci; Rokach; Shapira, 2015). However, successful applications in healthcare are relatively rare. In recognition of the importance of this domain and the opportunities it offers, our research community has recently organised a series of workshops to exchange ideas on the topic (Elsweiler *et al.*, 2016; Elsweiler *et al.*, 2017; Elsweiler *et al.*, 2018; Elsweiler *et al.*, 2019; Said *et al.*, 2020). Another indication of the interest of our community is the increasing number of secondary works dedicated to mapping and describing applications in this domain: at least 15 reviews on the application of recommender systems to healthcare have been published over the years, half of them in the last five years (Sezgin; Özkan, 2013; Kamran; Javed, 2015; Ferretto; Cervi; Marchi, 2017; Hors-Fraile *et al.*, 2018; Azmi; Abdullah; Emran, 2019a; Cheung *et al.*, 2019; Pincay; Terán; Portmann, 2019; Ertugrul; Elci, 2020; Su *et al.*, 2020; Croon *et al.*, 2021; Tran *et al.*, 2021; Cai *et al.*, 2022; Calderón-Blas *et al.*, 2023; Etemadi *et al.*, 2023; Vieira *et al.*, 2023).

In the following, we report the results of another review of the literature on healthcare recommender systems, with a focus on applications targeting older people. The need for this new review stems from the fact that the research questions approached by those reviews do not clarify three points of interest to this project, as described next.

3.1.1 Review planning

This review aims to establish the prevalence of healthcare recommender systems for older people that: (Q1) use data from standardised assessments of the target population, (Q2) provide explanations about provided recommendations, and (Q3) whether these explanations are faithful. The answers to these questions can show to what extent current blueprints and implementations of health recommender systems align with the way gerontologists assess their patients and create personalised care plans, as discussed in Section 2.1. The term "standardised assessment" is meant as an assessment in which the subject (the target user) is submitted to an instrument that has been systematically shown to be reliable, as discussed in Section 2.3. In the third question, the term "faithful explanation" denotes an explanation that faithfully reflects what the recommendation model actually computes. To answer these questions, the following protocol was adopted:

- Stage 1: Identification. Primary studies were identified by consulting three databases: Scopus, IEEEXplore, and the ACM/DL. The query submitted to the Scopus database follows the pattern "`system type AND domain AND population AND task`", with variables replaced by their corresponding expressions in Table 6. Since the same query submitted to the IEEEXplore and the ACM/DL databases did not return any results, we relaxed the search constraints to "`system type AND population`". Records from each study were gathered.
- Stage 2: Screening. Records that appeared multiple times were excluded, as well as studies whose abstracts indicated that the focus was not on healthcare recommender systems for older people; when in doubt, we consulted the full-text to assert the desired focus. When multiple reports of the same research project were found (e.g., CARE and +TV4E projects), one of the studies was selected for inclusion, at our discretion. Studies presenting early ideas about a healthcare recommender system that did not report results of a working prototype were included at our discretion.
- Stage 3: Collection. The full-text of the selected studies was gathered. Studies to which we did not have access through the licencing agreement between our university and the relevant publishers and were not available in the arxiv were excluded.
- Stage 4: Data extraction. The full text of each study was considered. We extracted evidence that the proposed recommender system uses data from standardised assessments of the target users to generate recommendations (Q1) and provides faithful explanations for the given recommendations (Q2 and Q3). Data about the adopted recommendation approach and the recruited participants were also collected.
- Stage 5: Inclusion of results from a previous review. The results of a review we conducted on the occasion of our qualification exam in January 2021 were included.

Table 6 – Variables used in searching for HRS studies targeting older adults

Variable	Expression
system type	"recommender system" OR "recommendation system" OR "recommending system" OR "recommender model" OR "recommendation model" OR "recommending model" OR "recommender method" OR "recommendation method" OR "recommending method" OR "recommender engine" OR "recommendation engine" OR "recommending engine" OR "recommender framework" OR "recommendation framework" OR "recommending framework" OR "recommender interface" OR "recommendation interface" OR "recommending interface" OR "recommender agents" OR "recommendation agents" OR "recommending agents" OR "collaborative filtering" OR "personalised recommendation"
domain	"health care" OR "healthcare" OR "wellness" OR "well-being" OR "wellbeing" OR "active ageing" OR "active aging" OR "healthy ageing" OR "healthy aging" OR "healthy lifestyle" OR "functional ability" OR "functional abilities" OR "intrinsic capacity" OR "assisted living"
population	"geriatrics" OR "geriatric" OR "gerontology" OR "gerontechnology" OR "gerontechnologies" OR "gerotechnology" OR "gerotechnologies" OR "gerontological" OR "elder" OR "elderly" OR "senior" OR "older person" OR "older adult" OR "older patient" OR "older people" OR "older population"
task	"intervention" OR "treatment" OR "assessment" OR "diagnostics"

Source: The author.

3.1.2 Review execution

The results of the execution are illustrated in Figure 9. In Stage 1, 122 records were identified: 56 records were retrieved from the Scopus database, 46 from the IEEEXplore, and 20 records from the ACM/DL. In Stage 2, 87 records were removed for varying reasons: 35 studies were excluded because they did not report on a recommender system (e.g., a study focused on some clinical recommendation framework); 27 were not focused on healthcare (e.g. educational needs of older people), 15 were early works (e.g. "initial ideas of a PhD trajectory"), five were duplicate, four were excluded because they were follow ups of a research project already covered, and one was not focused on older populations.

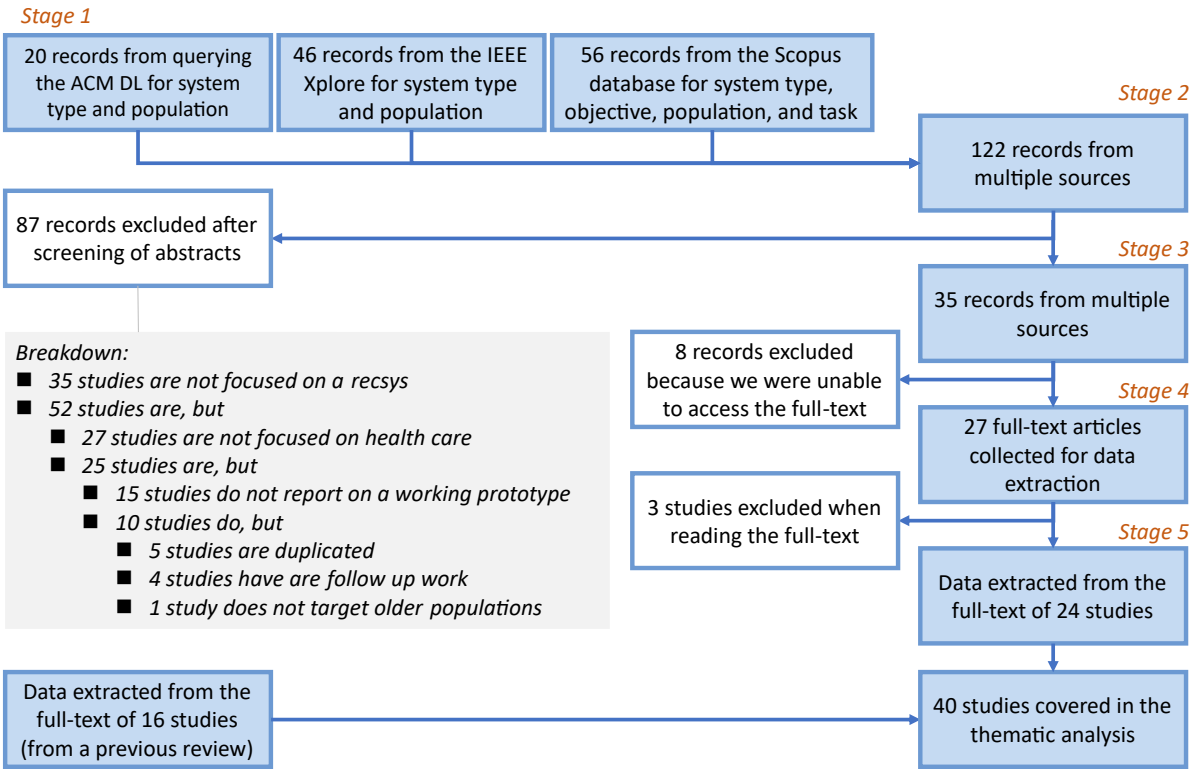
Of the 35 studies selected for data extraction (Stage 3), eight were excluded because we were unable to access the full-text of the study. Of the remaining 27, three studies were removed after closer reading. We extracted data from the remaining 24 studies and combined these results with data from a previous review based on a snowballing protocol. The data extracted from the final 40 studies are summarised in Table 7.

3.1.3 Review results

In summary, about one-third of the reviewed studies (13 out of 40) report using at least one standardised instrument or a structured questionnaire whose items are based on one of such instruments to collect data on the target user. Most of these instruments have been evaluated for their measurement properties from a psychometric perspective, as is the case of instruments used to measure intrinsic capacity and frailty[1]. In addition,

[1] It must be noted that evidence in support of the measurement properties of an instrument can accumulate from both psychometric and clinimetric perspectives, independently of how the underlying measurement model connects the latent variable to its indicators. Case in point: the Tilburg Frailty Indicator (Gobbens *et al.*, 2010; Gobbens *et al.*, 2012).

Figure 9 – Results obtained from the execution of the review protocol, by stage



Source: The author.

many studies use inventories that survey typical CGA domains, such as physical medical conditions, mental health conditions, daily functioning, and the social and environmental circumstances of an individual. It must be noted that some studies used instruments that were developed in the clinimetric tradition, such as the prognosis for patients with multiple pathologies (PROFUND) index and the clinical dementia rating scale (CDRS).

Regarding the provision of explanations, about one-sixth of the reviewed studies (6 out of 40) report a recommendation system that can generate explanations. Most of these systems provide explanations that are faithful (4 out of 6), and half of them have adopted an explanation style that consists in the visual display of quantitative information (3 out of 6), instead of the textual format that is typical of knowledge-based systems.

To organise a closer look, the reviewed studies were classified into five categories: (C1, with 2 studies) systems that use standardised assessment and provide faithful explanations, (C2, 11 studies) systems that use standardised assessment but do not provide explanations, (C3, 2 studies) systems that do not use standardised assessment but offer faithful explanations, (C4, 2 studies) systems that do not use standardised assessment but offer post-hoc explanations, and (C5, 23 studies) systems that do not use standardised assessment or provide explanations. No studies in which the system uses data from standardised assessments and provides post-hoc explanations were found.

The studies in category C1 are relatively recent. Lima *et al.* (2021) is an early contribution of our research project. A partner research group interviewed 108 older subjects for quality of life using the WHOQOL-BREF instrument. These assessments were combined with synthetic interventions to create the dataset used to train and evaluate the proposed recommendation model. In this model, both individual assessments and interventions are represented as polygons in a radar chart. The model decides to recommend an intervention based on the size of the intersection between the polygon representing the subject' scores and another polygon representing an intervention. This resembles a case-based approach in which the similarity between cases is computed using a geometric operation. The explanations, which are shown in a diagrammatic form similar to the one shown in Figure 1, are faithful in the sense that they reflect what the recommendation model actually computes. The expert-in-the-loop can inspect the size of each such intersection in the diagram and compare the resulting recommendation with her clinical judgement about the adequacy of the recommendation in face of the patient's assessment.

Román-Villarán *et al.* (2022) reports on a system that prescribes (or deprescribes) drugs to older patients with comorbidities and polypharmacy (i.e., patients that make regular use of five or more medications simultaneously to manage chronic diseases). This is a knowledge-based system that uses an ontology built on several clinical practice guidelines that have been published over the years. Similar to the previous study, both the recommendation and the explanation are directed to the attending professional in primary care, but this study differs in that both recommendations and explanations follow a textual style, as is typical in knowledge-based recommender systems. The authors report the use of several standardised assessments (both psychometric and clinimetric), but it is unclear how the system takes their scores into account when generating new recommendations.

The studies in category C2 make use of standardised assessments but do not provide any explanations for given recommendations. In most cases, the study used an instrument whose measurement properties have been established using evaluation procedures from the psychometric tradition (9 out of 11). Some examples of studies in this category are:

- Rist *et al.* (2015) adapted an instrument to assess well-being in older populations developed by the Stanford Center on Longevity (Kaneda; Lee; Pollard, 2011).
- Azmi, Abdullah and Emran (2019a) used data obtained from assessments performed by caregivers on nursing home residents. The assessment consisted of a battery of instruments for diverse domains, such as physical (e.g., health conditions, functional abilities, nutrition), cognitive, social, and environmental.
- Gannod *et al.* (2019) developed a recommender system to tailor the Preferences for Everyday Living Inventory instrument to residents in nursing homes (Haitsma *et al.*, 2013). The instrument captures the preference of a resident with regard to the

Table 7 – Results of the thematic analysis of the reviewed works

Study	Std. Assmt. / Measurand	System / Region	Exp. / Fthfl.	What is recommended to whom (and where/when and to what purpose)?
(C1) Systems that use standardised assessment and provide faithful explanations				
Lima *et al.* (2021)	Yes CGA	Poly Brazil (O)	Yes‡ Yes	health interventions to care professional during counselling with older patient in primary care
Román-Villarán *et al.* (2022)	Yes PROFUND	PITeS-TIiSS Spain (1)	Yes‡ Yes	drug prescription to care professional for older patients with polypharmacy in primary care
(C2) Systems that use standardised assessment but do not provide explanations				
Stiller, Roß and Ament (2010)	Yes CGA	Weitblick Germany (393)	No —	health, social, care, recreational, and household services close to older person's home
Rist *et al.* (2015)	Yes CGA	CARE EU (21)	No —	physical, mental, and social activities for older person living at home
Espín, Hurtado and Noguera (2016)	Yes malnutrition	NutElCare Spain (O)	No —	weekly, varied dietary plans to older persons living at home to prevent malnutrition
Vercelli *et al.* (2017)	Yes frailty	My-AHA EU (600)	No —	activities and lifestyle advice to older adults at home to prevent or manage frailty
Azmi, Abdullah and Emran (2019b)	Yes CGA	— Malaysia (139)	No‡ —	health interventions to improve well-being of older people living in nursing homes
Gannod *et al.* (2019)	Yes PELI	— USA (255)	No‡ —	suggestions to the nursing home manager to personalise services provided to residents
Zacharaki *et al.* (2020)	Yes frailty	FrailSafe EU (479)	No —	lifestyle and behavioural advice to older people at home to prevent frailty
Angelini *et al.* (2022)	Yes CGA	NESTORE EU (60)	No —	physical activities, social events, diet plans and games to older adults living at home
Frikha *et al.* (2024)	Yes IC	Senselife EU (O)	No —	Health and social care services, cultural and recreational events to older person at home
Llorente *et al.* (2024)	Yes dementia	PROCare4Life EU (93)	No —	guidance on using games for cognitive training to older people with PD/AD living at home
Kolakowski *et al.* (2025)	Yes IC	CAREUP EU (64)	No —	activities to older people at home to prevent decline in intrinsic capacity (IC)
(C3) Systems that do not use standardised assessment, but offer faithful explanations				
Robinson, Appiah and Yousaf (2017)	No —	— UK (10)	Yes Yes	lifestyle guidance to older people about how to reduce energy consumption at home
Rincon *et al.* (2019)	No —	EmIR† Portugal (E)	Yes Yes	health promoting activities to older person living at a nursing home
(C4) Systems that do not use standardised assessment, but offer post-hoc explanations				
Wu *et al.* (2023a)	No —	— USA (O)	Yes‡ No	health interventions to professionals in care of older people to prevent chronic diseases
Wu *et al.* (2023b)	No —	— USA (O)	Yes‡ No	health interventions to professionals in care of older people to prevent chronic diseases
(C5) Systems that do not use standardised assessment and do not provide explanations				
Luo, Tang and Thomas (2010)	No —	iPHR USA (?)	No —	home nursing procedural guidance and medical products to older patients living at home
Murakami *et al.* (2010)	No —	ApriPoco† Japan (?)	No —	exercise and lifestyle advice to older people at home to prevent declines in health
Bermingham *et al.* (2013)	No —	REMPAD Ireland (50)	No‡ —	video for groups of older people with dementia in reminiscence therapy at nursing home
Sasaki and Takama (2013)	No —	— Japan (5)	No —	outdoors, circular walking routes that are safe, compatible with an older person's locomotion

Results of the thematic analysis of the reviewed works (continued)

Study	Std. Assmt. / Measurand	System / Region	Exp. / Fthfl.	What is recommended to whom (and where/when and to what purpose)?
(C5) Systems that do not use standardised assessment and do not provide explanations				
Ponce *et al.* (2015)	No / —	QueFaire / Canada (E)	No / —	social events and transportation guidance for older people living at home
Silva *et al.* (2017)	No / —	+TV4E / Portugal (29)	No / —	videos on locally provided public services of interest to older person at home
Oliva-Felipe *et al.* (2018)	No / —	CaregiversPro / EU (1,200)	No‡ / —	psychosocial intervention for older person with dementia, educational resources to caregivers
Thakur and Han (2018)	No / —	— / UK (O)	No / —	activities to older people in smart homes to improve health
Besik and Alpaslan (2019)	No / —	RHCS / Türkiye (13)	No‡ / —	drug prescription to care professionals for geriatric patients in hospital setting
Martin, Costa and Cazorla (2019)	No / —	PHAROS† / Spain (20)	No / —	live guidance on how to perform exercises to older person at nursing home
Allalouf *et al.* (2020)	No / —	Tamaringa / Israel (24)	No / —	playlist for older person with dementia living at nursing home
Mishra, Busetty and Gudla (2020)	No / —	— / India (O)	No / —	activities to older people and caregivers in a smart home to assist with daily routines
Shinde *et al.* (2021)	No / —	IOTCARS / Germany (E)	No / —	lifestyle advice and reminders of care routine to older people at home to improve health
Sitparoopan *et al.* (2021)	No / —	— / Sri Lanka (O)	No / —	recipes to elderly people at home to manage malnutrition and obesity
Dhivakar (2022)	No / —	— / India (O)	No / —	advice to older patients at home after hospital discharge to provide continued care
Minakata *et al.* (2022)	No / —	— / Japan (O)	No / —	outdoors walking routes with low tripping risk, compatible with an older person's locomotion
Martín *et al.* (2023)	No / —	DigiHEALTH / EU (60)	No / —	habit modification guidance to older adults at home to promote healthy ageing
Huang and Palaoag (2024)	No / —	— / China (O)	No / —	care services to older people and family at home to improve quality of life
Prasetyo and Baizal (2024)	No / —	— / Indonesia (16)	No / —	recipes to older people at home to meet balanced nutritional needs
Yuan, Zhang and Zhou (2024)	No / —	— / China (O)	No / —	health-related short videos to older people at home
Zhou (2024)	No / —	— / China (O)	No‡ / —	recipes to nursing home manager to balance nutrition and taste of served meals
Bentlage *et al.* (2025)	No / —	PROCare4Life / EU (316)	No‡ / —	physical activity to older people with Parkinson disease at home, daycare or rehab centre
Jiang (2025)	No / —	— / China (O)	No / —	health education resources to older people at home to improve self-care ability

Source: The author.

Note: Studies are listed in chronological order within each category. The 2nd column indicates if/which standardised instruments are used. The 3rd column shows the system name and the region in which the study participants reside. A dagger beside the system name denotes a human-robot interface. The number beside the region is the number of participants; (O) indicates an offline evaluation, and (E) that this is an early study which does not report results. The 4th column shows if explanations are provided, and if they are faithful; a double dagger indicates an expert-in-the-loop. Topic of recommendation is highlighted in the 5th column.

services that are provided by the institution (e.g., the resident wants to choose her clothes). The aim is to shorten the interview because the instrument has 72 questions, which makes its application difficult given the daily duties of the caregivers.

- Although Stiller, Roß and Ament (2010) and Orte *et al.* (2018) did not identify the use of any particular instrument, the authors reported the use of questionnaires to assess a person in all the domains that are usually surveyed in CGAs, such as physical, psychological, social, and environmental domains.

A common feature of these studies is that the recommendation is presented directly to the user (9 out of 11), with the exception of two studies in which the recommendations are directed at the nursing home staff (Gannod *et al.*, 2019; Azmi; Abdullah; Emran, 2019b). Consequently, the topic of the recommendation falls predominantly in one of these classes: information about health, social, and recreational services available in the neighbourhood of the user, activities for the user to do at home (e.g., physical exercises to maintain or improve health, playing computer games for cognitive improvement), and lifestyle advice (diet plans, guidance to adapt home to avoid adverse events).

The studies in categories C3 and C4 report on systems that do not use standardised assessments, but do provide explanations. The two studies in C2 are knowledge-based systems that provide faithful explanations. The system proposed by Robinson, Appiah and Yousaf (2017) recommends adaptations to the daily routine in order to reduce energy consumption at home, and the explanation given is based on the amount of money that the system estimates will be saved by adopting the recommendation, while the system designed by Rincon *et al.* (2019) recommends physical activities for older people in nursing homes, and also explains the recommendation in terms of the expected benefit, such as "because your doctor has recommended it for lowering your cholesterol". In both cases, the explanations simultaneously represent the "logic" behind the recommendation and are phrased so that the benefit of following the recommendation is explicit. In contrast, the two studies in the category C4 provide post-hoc explanations. They correspond to reports by the same research group of attempts to provide explanations to black-box models applied to diagnose chronic diseases that are highly prevalent in older populations. The prediction is based on data from patient medical records including test reports, treatment histories, diagnostic records, and other health-related features (Wu *et al.*, 2023a; Wu *et al.*, 2023b). Although these are post-hoc explanations, they resemble the approach in Lima *et al.* (2021) in that explanations consist of the visual display of quantitative data representing the relative importance of specific features of the patient's data in the assignment of one or more labels to a new patient (e.g., has-diabetes).

Finally, the majority of the reviewed studies fall into category C5 (23 of the 40): studies reporting on recommender systems that do not use standardised assessments and do

not provide explanations. Similar to the studies in category C2, the studies in this category report on recommender systems in which the recommendation is presented directly to the target user (19 of the 23), with the exception of a few studies in which the recommendation is shown to a care professional (Bermingham *et al.*, 2013; Oliva-Felipe *et al.*, 2018; Besik; Alpaslan, 2019; Zhou, 2024). The topic of the recommendations that are directed at the target user falls predominantly in one of these classes: music or video for educational or therapeutic purposes, safe outdoor walking routes, information about health, social, and recreational services available in the neighbourhood of the user, activities for the user to do at home, and lifestyle advice (diet plans, guidance to adapt home to avoid adverse events). In contrast, the topics of the recommendations that are shown to the care professional are, for example, recipes to balance nutritional value and tastes of residents in a nursing home (Zhou, 2024), videos to be used in a group reminiscence therapy for older people with dementia (Bermingham *et al.*, 2013), as well as other psychosocial interventions for older people with dementia (Oliva-Felipe *et al.*, 2018). The study by Besik and Alpaslan (2019) is an outlier in that their system recommends to the care professional the prescription of drugs for an older patient, as does the system proposed by Román-Villarán *et al.* (2022) in category C1, but it does not seem to provide explanations for the recommendations given.

3.1.4 Conclusion

The results just presented show that most health recommender systems that target users from older populations do not use data from standardised instruments to produce recommendations (i.e., 27 out of 40 studies were negative for Q1). Moreover, the majority of the studies we surveyed report on systems that do not have explanation facilities (34 out of 40 were negative for Q2), and among the recommender systems that do provide explanations, only two-thirds generate faithful explanations (4 out of 6). This seems to be in agreement with the assumption that, in most cases, the object of the recommendation does not have the potential to produce significantly harmful outcomes: the recommendations are presented directly to the target user, who has the adequate experience to judge their relevance and decide to follow the recommendation only when appropriate.

The results also show a disconnect between current blueprints and implementations of healthcare recommender systems for older adults and the practice in gerontological primary care described in Section 2.1. This is because professionals in the care of older people frequently rely on standardised instruments to perform comprehensive patient assessments, and use the outcome of these assessments to develop personalised care plans for the patients. In addition, the introduction of an expert-in-the-loop to mitigate risks in a high-stakes scenario would require the supporting recommender system to provide faithful (and interpretable) explanations, so that the care professional can understand the rationale behind the recommendation and assess its merits and shortcomings.

3.2 Multilabel classification and label ranking tasks

The idea of using computers to support clinical research and practice is not new. Pioneering work on computer-aided medical diagnosis, including emerging applications and seminal discussions, began as early as the late 1950s (Ledley, 1960). Several original computer systems that were "prepared" to showcase the methods that were being introduced can be understood as precursors to contemporary neighbourhood-based and Bayesian models for classification tasks (Tanimoto, 1960; Collen *et al.*, 1964). Since then, computer systems and computational methods have evolved dramatically, but those early tasks remain a significant research topic, as recent applications of machine learning demonstrate. For example, Kavuluru, Rios and Lu (2015) report on the performance of multiple machine learning models for the automatic assignment of standardised diagnosis codes to electronic medical records. In a similar vein, Yang, Shi and Ni (2021) report on the creation of a public collection of lightweight medical image datasets that they then use to assess the performance of several models in solving a diverse set of classification tasks.

In a nutshell, multilabel classification refers to the task of learning how to map an instance to a set of predefined labels (Brinker; Fürnkranz; Hüllermeier, 2006). For example, the task of predicting what kind of food (Arabic, Brazilian, Chinese, etc.) a person is inclined to consume in food trucks is commercially valuable, and learning how to do it can be framed as a multilabel classification task (Rivolli; Parker; Carvalho, 2017). Label ranking generalises this task, as it requires not only mapping instances to sets of labels, but also sorting the labels based on some measure. In the previous example, the ranking could tell us that a person likes Italian food the best, and prefers Chinese to Greek dishes.

In this section, we review how multilabel classification and label ranking tasks are formally specified. These definitions are then used to show that the task of creating personalised care plans can be modelled as either of these tasks. This is followed by a brief discussion about the predominant approaches to solving each task, the metrics that are commonly used to evaluate different dimensions of model performance, and the benchmark datasets that the research communities use to assess the merits of new models.

3.2.1 Notation, concepts, definitions, and data structures

Let the tuple $D := (X, Y)$ represent a dataset, with X being an (m, d)-matrix holding the data of m subjects regarding d attributes, and Y being an (m, n)-matrix holding the data about how each subject is associated with each of the n labels in the label set $\mathcal{L}$. By convention, let's call X the description matrix, and Y the assignment matrix.

Let $X_i = (x_{i0}, \ldots, x_{i,d-1}) \in \mathcal{X}$ denote the i-th row of X, with x_{ik} representing the score the i-th individual obtained for the k-th attribute, and $\mathcal{X}$ is the example space, broadly conceived (i.e., x_{ik} can take a categorical or numeric value, and no value is missing).

In the context of multilabel classification, we say that Y is a multilabel assignment. We adopt the usual encoding in benchmark datasets, in which $y_{ij} = 1$ indicates that the i-th individual is assigned to the j-th label, and $y_{ij} = 0$ indicates the opposite[2]. In this setup, the solution of a multilabel classification problem is a function $h : \mathcal{X} \to 2^{\mathcal{L}}$, which is often called a hypothesis to highlight that $h \in \mathcal{H}$, the hypothesis space that is spanned by a model. Following Madjarov *et al.* (2012), let q be a quality criterion that rewards hypotheses with high predictive performance and low complexity. Then, the aim of a multilabel classification task is to learn a function $h \in \mathcal{H}$ that maximises $q(h, D)$.

Unlike multilabel classification problems, in which a solution is a subset of $\mathcal{L}$, the structure of a solution for a label ranking problem depends on the variant of the problem at hand. In some settings, ties are allowed, which means that two or more labels can hold the same position in a ranking. This variant is not within the scope of our project, but interested readers can find detailed discussions elsewhere (Gionis *et al.*, 2006; Jiménez, 2023). This review focuses on variants that can be formalised as rankings with no ties.

To that purpose, let's introduce the relation $\succ$, which represents the preference among two labels: $a \succ b$ means that a is preferred to b (or that a precedes b in a ranking). For a label ranking assignment Y, we adopt the usual encoding in benchmark datasets. For example, assume $\mathcal{L} = \{0, 1, 2, 3\}$. Then, $Y_i = (1, 2, 0, 3)$ means that the i-th individual is associated with the ranking $1 \succ 2 \succ 0 \succ 3$. This is a complete ranking because all labels in $\mathcal{L}$ appear in the assignment for the i-th subject. To encode incomplete rankings, we use -1 as a filler: $Y_i = (1, 2, -1, -1)$ encodes the ranking $1 \succ 2$. In this setup, the solution of a label ranking problem (with no ties and that allows for incomplete rankings) is a function $h : \mathcal{X} \to \mathscr{A}(\mathcal{L})$, with $\mathscr{A}(\mathcal{L})$ being the union of the permutations of all subsets of $\mathcal{L}$:

$$\mathscr{A}(\mathcal{L}) = \bigcup_{s \in 2^{\mathcal{L}}} \Pi(s). \tag{3.1}$$

This is not a standard presentation. In the literature, formulations of the label ranking problem from an order-theoretic perspective are the norm (i.e., they are built from preference relations), and settings in which the rankings have a fixed size are predominant (Cheng; Hühn; Hüllermeier, 2009; Fürnkranz; Hüllermeier, 2011; Fotakis; Kalavasis; Psaroudaki, 2022; Jiménez, 2023). We adapted the definition of "multilabel ranking" from Brinker, Fürnkranz and Hüllermeier (2006) by giving it a combinatorial expression. This adaptation was made to highlight the fact that rankings of varied sizes can be encoded in the same assignment Y, e.g., given two subjects a and b, it may be the case that $Y_a = (1, 2, 0, 3)$ and $Y_b = (1, 2, -1, -1)$. The reason this is important is discussed later, but from now on the term "label ranking problem" refers to the problem defined above. As before, the aim of a label ranking task is to learn a function $h \in \mathcal{H}$ that maximises $q(h, D)$.

[2] Since $\mathcal{L}$ is a set, the term j-th label is meaningless. To remediate this issue, we ask the reader to interpret $\mathcal{L}$ as a set or sequence, depending on the context. For example, $\mathcal{L} = \{0, \ldots, n-1\}$ in a combinatorial context such as $2^{\mathcal{L}}$, or $\mathcal{L} = (0, \ldots, n-1)$ when order is needed for meaning.

As seen, the assignment matrix plays a defining role in both problems. Often, researchers face the need to describe a dataset to a learned audience. To do this, they resort to the metrics in the literature that have been devised to describe certain aspects of an assignment matrix (Herrera *et al.*, 2016). For example, the cardinality metric refers to the average number of labels per instance and, considering the specific encodings for multilabel and label ranking assignments described earlier, it can be computed as follows:

$$\mathrm{Card}(Y) := \frac{1}{m}\sum_i \#Y_i, \text{ with } \begin{cases} \#Y_i := \sum_j [\![y_{ij} > 0]\!] & \text{for multilabel assignments,} \\ \#Y_i := \sum_j [\![y_{ij} \geq 0]\!] & \text{for label ranking assignments,} \end{cases}$$

with $[\![P]\!] = 1$ if the statement P is true, and zero otherwise. In the formal setup described earlier, the range of this metric is the $[0, n]$ interval. For multiclass classification problems, which are the particular case of multilabel classification in which every instance is assigned to a single label, the cardinality is one. For label ranking problems that require complete rankings, the cardinality is n. A normalised version of this metric, known as density, is computed by dividing the cardinality of Y by n. The density metric can be understood as the average number of cells of a matrix Y that actually map a subject to a label.

It is clear that cardinality roughly describes the rows of an assignment matrix, and density focus on its cells. There are many metrics that can play a similar role for columns of an assignment matrix. For example, the maximum imbalance ratio computes the ratio between the most and least frequent labels. For a multilabel assignment Y, it computes:

$$\mathrm{MaxIR}(Y) := \frac{\max_{j\in\mathcal{L}} \sum_i Y_{ij}}{\min_{j\in\mathcal{L}} \sum_i Y_{ij}},$$

where $\sum_i Y_{ij}$ is the frequency of the j-th label (Herrera *et al.*, 2016). The range of this metric is the $[1, m)$ interval[3]. The lower bound represents a perfectly balanced assignment (i.e., all labels occur with the same frequency), and the higher the value, the more severe the imbalance. The fact that this metric depends on the size m of the dataset can be overcome by taking advantage of the complement of its reciprocal, whose range is $[0, 1)$:

$$\mathrm{Imb}(Y) := 1 - \frac{\min_{j\in\mathcal{L}} \sum_i y_{ij}}{\max_{j\in\mathcal{L}} \sum_i y_{ij}}. \tag{3.2}$$

For a label ranking assignment Y, it is convenient to "downgrade" it to a multilabel assignment $Y^{\downarrow}$ and then compute the metric $\mathrm{Imb}(Y^{\downarrow})$. The downgrade operator computes:

$$Y^{\downarrow} := (y^{\downarrow}_{ij})_{m\times n} \text{ such that } y^{\downarrow}_{ij} = \begin{cases} 1, & \text{if } j \in Y_i, \\ 0, & \text{otherwise,} \end{cases} \tag{3.3}$$

e.g., if we have an assignment $Y_b = (1, 2, -1, -1)$, then it is the case that $Y^{\downarrow}_b = (0, 1, 1, 0)$.

[3] Assuming that the degenerate case in which a label can have frequency zero is excluded.

Two other recurring metrics in the literature, which involve the concepts of labelset and single labelset, focus on a combinatorial characterisation of an assignment. The term "labelset" (not to be confused with the label set $\mathcal{L}$) refers to a combination of labels that appear in an assignment Y. In other words, the number of labelsets of an assignment Y corresponds to the number of distinct rows of the matrix Y. To illustrate this idea, let's take the Iris dataset as an example. It has $m = 150$ instances and $n = 3$ labels. In its multiclass version, in which each instance is assigned to a single label, there are three labelsets, namely $(1, 0, 0), (0, 1, 0)$ and $(0, 0, 1)$. As we shall soon see, there is a version of this dataset in the Paderborn repository, which is a collection of datasets used for benchmarking label ranking models (Cheng; Hühn; Hüllermeier, 2009). In this version, each instance is assigned to a complete ranking. In principle, there could be $n! = 6$ labelsets, but one of them never occurs, namely $(0, 2, 1)$. The labelsets that do occur appear in multiplicity, e.g., the labelset $(0, 1, 2)$ appears 50 times. The term "single labelset" refers to a labelset that appears only once in an assignment. For example, the foodtruck dataset that comprises the Cometa repository for multilabel classification benchmarking has $m = 407$ instances and $n = 12$ labels. Of the $n! = 479,001,600$ possible labelsets, only 117 were observed in the assignment Y, and 74 are single labelsets (Rivolli; Parker; Carvalho, 2017).

There are many other useful metrics in the literature (Herrera *et al.*, 2016), but we chose to present these because they concisely summarise the distribution of labels in an assignment matrix from distinct but complementary perspectives: rows, columns, and cells, as well as a contrast between potential and observed arrangements of labels.

3.2.2 The task of creating personalised care plans

Earlier we mentioned a study reported by Tavassoli *et al.* (2022) in which her research group collected the assessment of intrinsic capacity of more than ten thousand older participants. After a screening process, a team of primary care professionals met with the participants with health issues, and a personalised care plan was created for each of these individuals. The findings of the study indicate that a set of 22 different interventions or referrals were made by attending professionals (Tavassoli *et al.*, 2022, see Table 3).

From a computational perspective, one could frame the creation of care plans based on patient assessment data as a multilabel classification problem: The set of labels $\mathcal{L}$ is the set of interventions or referrals that the regional health system provided. The patient assessments provide the description of the subjects in X, and the recommendations made by the group of attending professionals correspond to the assignment matrix Y. The task corresponds to learning a function h that maps a patient assessment X_i to a care plan $\hat{Y}_i$ that contains exactly the same recommendations made by the attending professional. In other words, the success of this learning task is measured by the degree to which h can replicate the decision-making of the care professionals in this particular study setting.

Similarly, if the order in which an intervention or referral appears in a care plan is important, a new constraint must be introduced. This constraint, although more challenging to satisfy, may be useful as it can orient the patient and family on how to prioritise their available resources to obtain the best health outcome for the patient. For example, in the CGA template promoted by the SBGG that we mentioned in Section 2.1, there is a section in which the care professional records the personalised care plan that was agreed with the patient and her family. In this section, recommended interventions are separated by priority, with interventions that are capable of reducing functional decline having the highest priority. From a computational modelling perspective, this could be framed as a label ranking task as we defined earlier: solutions are incomplete rankings with no ties.

3.2.3 Solving multilabel classification tasks

In the literature, methods for solving multilabel classification tasks are traditionally divided into two broad categories: problem transformation and algorithm adaptation approaches[4] (Tsoumakas; Katakis, 2007; Madjarov *et al.*, 2012; Herrera *et al.*, 2016).

In the problem transformation approach, one of the simplest methods consists in decomposing a multilabel assignment Y into n single-label assignments (one for each original label in $\mathcal{L}$), which are then treated as n independent binary classification problems specified as $D_j = (X, Y_{:j})$ for $j \in 0 \ldots (n-1)$. Accordingly, n separate solutions $h_j : \mathcal{X} \to \{0, 1\}$ are learned and then concatenated to produce a solution for the original multilabel problem, namely $h = (h_0, \ldots, h_{n-1})$. Thus, in prediction time, $\hat{Y}_i = h(X_i) = (h_0, \ldots, h_{n-1})(X_i)$. Any model that supports binary classification tasks can be a "template" for h_j, such as logistic regression or decision trees. In the literature, this is known as the binary relevance method, which "could be considered as a simple ensemble of binary classifiers with a very straightforward strategy to fuse their individual predictions" (Herrera *et al.*, 2016).

Regarding the algorithm adaptation approach, the idea is to choose a model that supports binary classification in single output, and adapt its inner workings to produce multiple outputs. More precisely, the aim is to adapt a model that generates hypotheses $h \in \mathcal{H}$ such that $h : \mathcal{X} \to \{0, 1\}$ to the multilabel setting. For example, the C4.5 algorithm was devised to induce decision trees from data. In its original conception, each leaf of a tree stores a single value, which represents the label to be assigned to an instance whose evaluation leads up to that particular leaf. The ML-C4.5 is an adaptation of this algorithm for handling multiple outputs. The adaptation consists not only of extending the content of the leaf nodes (to store multiple labels), but also of replacing the node-splitting entropy measure by a function that considers multiple labels at once (Herrera *et al.*, 2016).

[4] In some sources, the problem transformation approach is sometimes referred to as the data transformation approach, for reasons that will soon become clear. Similarly, the algorithm adaptation approach is sometimes referred to as the method adaptation approach.

In a comprehensive study that compares the performance of multilabel classifiers, with plenty of representatives of both approaches described above, Bogatinovski *et al.* (2022) report that two models, one from each approach, appear at the top of the performance ranking. These models, identified in their study as Binary Relevance with Random Forests of Decision Trees (RFDTBR) and Random Forest of Predicting Clustering Trees (RFPCT), are both based on random forests[5]. The study compared 26 methods on 42 datasets across a spectrum of 18 evaluation metrics. The datasets used in the evaluation have been made publicly available by the curators of the Cometa repository (Charte *et al.*, 2018).

Regarding the metrics used to evaluate performance, a common categorisation in the literature divides them into example- and label-based metrics. In the following, we describe four of the metrics used in Bogatinovski *et al.* (2022), beginning with the subset-accuracy. This example-based metric corresponds to the fraction of predicted labelsets, $\hat{Y}_i$ for $i \in \mathcal{I}$, that exactly match their respective ground truth. More precisely,

$$\text{subset-accuracy}(Y, \hat{Y}, \mathcal{I}) := \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} [\![Y_i = \hat{Y}_i]\!],$$

where $\mathcal{I}$ is the set of indices of the test instances, and $[\![P]\!]$ is the Iverson bracket seen earlier. The Hamming loss, which is also an example-based metric, corresponds to the fraction of predicted cells in $\hat{Y}$ that diverge from their respective ground truth:

$$\text{hamming-loss}(Y, \hat{Y}, \mathcal{I}) := \frac{1}{|\mathcal{I} \times \mathcal{L}|} \sum_{(i,j) \in \mathcal{I} \times \mathcal{L}} [\![y_{ij} \neq \hat{y}_{ij}]\!].$$

It must be noted that both metrics range in the $[0, 1]$ interval. From a certain perspective, example-based metrics describe a tensor $\mathbf{Y} = (Y \otimes e_0 + \hat{Y} \otimes e_1)$ by its horizontal slices[6], resembling the mechanics behind the cardinality and density metrics. Although the subset-accuracy measures successes and the Hamming loss measures failures, note that they are not necessarily complementary. Of course, when subset-accuracy is one, then it must be the case that the Hamming loss is zero, and vice-versa. However, both can be simultaneously very low, indicating severe error dispersion. For example, let $D = (X, Y)$ be a multilabel dataset with $n > 2$ labels, in which each test instance is assigned exactly two labels. Suppose that, for each test instance, the model commits a single error, which is always a false positive. This setting implies a subset-accuracy of zero, and the Hamming loss is $1/n$.

In contrast, label-based metrics describe the tensor $\mathbf{Y}$ by its lateral slices. Examples of label-based metrics that were used in the evaluation reported by Bogatinovski *et al.* (2022) are the macro- and micro-averaged F1 scores, both based on the standard F-measure:

$$\text{F-measure}(tp, fn, fp) := \frac{2\,tp}{fn + fp + 2\,tp}, \tag{3.4}$$

[5] We highlight these models because they will be used in Chapter 5 as baselines, and their structural similarity with our model will be used to discover areas for potential improvement.

[6] This corresponds to the 3-tensor $\mathbf{Y}$ whose frontal slices are $\mathbf{Y}_{::0} = Y$ and $\mathbf{Y}_{::1} = \hat{Y}$.

with tp as the number of true positive cases, fn is the number of false negative cases, and fp is the number of false positive cases (Rivolli *et al.*, 2020). In the macro-averaged F1 score, these variables are computed for a lateral slice $\mathbf{Y}_{\mathcal{I}j}$ and then "plugged" into the Equation 3.4; estimates for the F-measure are collected for each slice and then averaged. A variant of this metric, the label-weighted F1 score, differs only in the averaging rule: the frequency of each label in Y is used to weight the estimates from each lateral slice. Finally, in the micro-averaged F1 score, these variables are computed globally in the $\mathbf{Y}_{\mathcal{I}}$ tensor and then "plugged" into the Equation 3.4. The range of these metrics is the $[0, 1]$ interval.

3.2.4 Solving label ranking tasks

As seen in Section 3.2.2, the problem of building a personalised care plan can be modelled as a label ranking task that handles a modest number of labels (compared to traditional applications of recommender systems). This makes the literature on preference learning a natural field for looking for solutions (Fürnkranz; Hüllermeier, 2011). In this tradition, one of the most versatile approaches focuses on learning a separate scoring (or utility) function for each individual label, which can then be aggregated in a ranker.

The fundamental idea of this approach is to learn a set of "building block" functions, namely $g_j : \mathcal{X} \to \mathbb{R}$ for $j \in 0 \ldots n-1$, whose outputs can be combined into a valid ranking. More precisely, let $D = (X, Y)$ be a label ranking dataset as described in Section 3.2.1, and define $g(X_i) := (g_0, \ldots, g_{n-1})(X_i)$ as a concatenated scoring function. Then, a naive solution is $\succ g(X_i) := \arg\operatorname{sort} -g(X_i)$ (Fotakis; Kalavasis; Psaroudaki, 2022). For example, let $\mathcal{L} = \{0, 1, 2\}$ and assume that $g(X_i) = (0.2, 0.4, 0.8)$ for some instance X_i. Then the following ranking is obtained: $\succ g(X_i) = \arg\operatorname{sort}(-0.2, -0.4, -0.8) = (2, 1, 0)$.

However, this is not a particularly good solution to handle incomplete rankings. A common remedy to this issue is to combine a multilabel classifier and a label ranker so that irrelevant labels are removed from the ranker's output (Brinker; Hüllermeier, 2007). Building on the previous example, let $g^{\downarrow} : \mathcal{X} \to 2^{\mathcal{L}}$ be a solution to the multilabel classification problem $(X, Y^{\downarrow})$, and assume that $g^{\downarrow}(X_i) = (0, 1, 1)$, which corresponds to the label subset $\{1, 2\}$. Then, the output of the combined ranker-classifier h should be:

$$h(X_i) := (\succ g \cap g^{\downarrow})(X_i) = \succ g(X_i) \cap g^{\downarrow}(X_i) = (2, 1, 0) \cap \{1, 2\} = (2, 1),$$

where $\cap$ represents an asymmetric relation between a sequence and a set, such that the elements of the sequence that are not in the set are "erased" from the sequence.

Another popular approach to label ranking problems is based on decision trees. Similar to the algorithm adaptation approach seen in multilabel classification, the idea is to adapt an algorithm that induces decision trees to represent rankings in the leaves, and modifying the node splitting function to make an efficient use of the ranking data. In fact, Zhou and Qiu (2018) find that tree-based models achieve highly competitive results

compared to models following other traditional approaches to label ranking, and more recently, Fotakis, Kalavasis and Psaroudaki (2022) report that random forests usually rank higher than other tree-based models. The first study compared the performance of an adapted random forest against 11 other models on 16 datasets, and the second study reports a comparison of an adapted random forest against four other tree-based models on 21 datasets (16 semi-synthetic, 5 real-world). Both studies used the datasets that have been made publicly available by the curators of the Paderborn repository (Cheng; Hühn; Hüllermeier, 2009). Regarding the performance metric, both studies report results using a variant of the Kendall's τ coefficient for tie-free rankings, averaged over the test cases:

$$d_\pi(j, l, Y_i) := \begin{cases} \pi(l, Y_i) - \pi(j, Y_i) & \text{if } j, l \in Y_i, \\ 0 & \text{otherwise,} \end{cases}$$

$$d_\tau(Y_i, \hat{Y}_i) := \sum_{j<l} [\![\, d_\pi(j, l, Y_i)\, d_\pi(j, l, \hat{Y}_i) < 0 \,]\!],$$

$$\tau_a(Y_i, \hat{Y}_i) := \begin{cases} 1 - \dfrac{4 d_\tau(Y_i, \hat{Y}_i)}{n(n-1)} & \text{if } |Y_i \cap \hat{Y}_i| > 1, \\ \text{undefined} & \text{otherwise,} \end{cases}$$

$$\text{ktau}(Y, \hat{Y}, \mathcal{I}) := \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \tau_a(Y_i, \hat{Y}_i), \tag{3.5}$$

where $\pi(j, Y_i)$ is the position of the j-th label in Y_i, e.g., if $Y_i = (2, 1, 0)$, then $\pi(0, Y_i) = 2$.

As a final thought, recall from the previous section that two of the top-ranking models in the extensive evaluation reported by Bogatinovski *et al.* (2022) were also based on random forests. Because of its success in both tasks, it will be explored during the evaluation of our model. For this reason, we must highlight some details about the architecture of random forests. It follows an ensemble strategy to induce multiple decision trees during training. Each tree is induced from a random subset of the training data and, at each node, only a random subset of features is considered for splitting. This double-bagging strategy improves the bias-variance trade-off of decision trees that are allowed to grow unpruned. In analogy to the way the size of multilayer perceptrons is measured, the size of a random forest is computed by summing up the sizes of its individual trees, and the size of a tree is calculated by summing up the number of "weights" needed to represent its internal and terminal nodes. For example, if the trees are induced using the CART algorithm, then the resulting trees are binary trees, and each decision node can be represented with two weights (one to indicate a feature, and one to hold its corresponding threshold value), and each leaf can be represented by a single weight (that indicates a labelset). The fact that the number of trees can be specified endows the random forest with the ability to produce models with an arbitrarily large size. Finally, in prediction time, the model merges the results of the individual trees (e.g., by majority vote in classification, by averaging their outputs for regression) to produce a combined prediction.

3.3 Task-based effectiveness of elementary visualisations

Diagrams are part of our everyday experience: they appear in daily weather forecasts, financial reports in banking apps, and smartwatch apps that estimate how many steps you took this week. They play a key role in the communication of results in academic papers. In the history of engineering, the Smith chart, which was used as a calculating device by electrical engineers before the spread of electronic calculators, gained a physical analogue with a rotating rule (Maddio; Pelosi; Selleri, 2023). Looking further back in history, we find the innovative use of diagrams as a dialogical element in the axiomatic method of Euclid (Digregorio, 2020; Dutilh-Novaes, 2025). Yet, despite this presence, they resist definition.

In philosophy and history of logic, researchers resort to the concept of iconic sign in semiotics as a first but admittedly precarious approximation: diagrams "are signs which resemble what they signify ... by mimicking the structures they signify", as Legg (2013) argues[7]. Closer to home, the visualisation research community defers this issue by delimiting the scope of interest to visual displays of data that are relevant for and interpretable by users performing certain tasks. These tasks are then placed at the centre of their research agenda. The latter point is demonstrated by the large number of diverse and systematic classification schemes (e.g., typologies and taxonomies) for visualisation-supported tasks that the community has developed over the years (Dimara; Stasko, 2022). In this review, we focus on tasks in which the user is prompted to make a decision, and thus the role of the diagram is to facilitate the successful completion of a decision-making task.

As stated by Dimara and Stasko (2022), most classifications omit decision-making tasks. Instead, they usually focus on analytic tasks, which are tasks in which the user's goal is served by querying a dataset. An exception is the typology devised by Brumar *et al.* (2025), which focuses on decision-making tasks and classifies them into three composable categories: choose, activate, and create. Given an initial set of options as input, the first corresponds to selecting the top k options by comparing them with each other; the second to independently evaluating the options against a threshold; and the third to combining or generating new options or new information. Being conceptual by design, this typology does not indicate which basic visualisations, such as bar charts or scatterplots, are more effective in a given decision-making scenario. As we shall soon see, this guidance is available for low-level analytic tasks, such as those in the taxonomy created by Amar, Eagan and Stasko (2005). Among the categories in the latter, one finds (a) the sort tasks, whose goal is "given a set of data cases, [to] rank them according to some ordinal metric", (b) the find extrema tasks, aiming at "find[ing] data cases possessing an extreme value of an attribute over its range within the dataset", and (c) the filter tasks, whose goal is "given some concrete conditions on attribute values, [to] find data cases satisfying those conditions."

[7] She concludes that a diagram is a sign that can have iconic, symbolic (e.g., the numbers in a scale) and indexical components (e.g., the prime centre of a Smith chart). Examples are mine.

In a classification scheme, items are grouped into categories by their similarity with respect to some attribute(s). The attributes and similarity criteria are selected based on the fruitfulness of the regularities known by the researchers to hold in their domain (Hoyningen-Huene, 2008). In our view, items are grouped by structural similarity in the typology by Brumar *et al.* (2025), as the salient attributes that separate the classes are the number and structure of options produced by the task. In contrast, items are grouped by functional similarity in the taxonomy by Amar, Eagan and Stasko (2005); in other words, the user's goal when performing a task is the salient feature that separates the classes.

Based on this complementarity, one could argue that, in a decision-making scenario in which the user must inspect a dataset, there is a correspondence between these schemes: (a) the inputs and outputs of the concatenation of a sort task and a find extrema task is described by the choose task, and (b) the activate task describes the inputs and outputs of a filter task. This circumstantial correspondence between decision-making and analytic tasks can be evoked to extrapolate the empirical results in the literature to attempt an answer to the question about the effectiveness of basic visualisations in decision-making tasks. For example, Saket, Endert and Demiralp (2019) conducted a user study to assess five visualisations (bar chart, pie chart, line chart, scatterplot, and tabular presentation) across the ten low-level analytic tasks in the taxonomy by Amar, Eagan and Stasko (2005). Each participant was randomly assigned to a category (e.g., filter) and, in the main part of the study, was asked to complete 30 tasks (five visualisations, two datasets, and three trials). The authors report that the bar chart is strongly associated with higher participant performance in general, and associated with highly competitive participant performance in the analytic tasks we singled out earlier, which we associated with decision-making tasks, as summarised in Table 8.

Table 8 – Effectiveness of basic visualisations on three analytic tasks

	Visualisation-supported low-level analytic tasks										
	sort			find extrema			filter				
Diagram	acc	ct	pref	acc	ct	pref	acc	ct	pref	sc1	sc2
bar chart	1	1	1	1	1	1	1	1	1	9	6
scatterplot	1	1	2	1	1	2	1	2	2	5	3
tabular	1	2	2	1	2	2	1	1	1	5	4

Source: The author.

Note: Cross-section of the results reported by Saket, Endert and Demiralp (2019) for the analytic tasks we singled out in the text. Here, "acc" stands for accuracy, "ct" for completion time, and "pref" for explicit participant preference. The numbers in the columns represent the rank of a visualisation, as reported in Table 1 of that study ("1" means top performance). The column "sc1" shows the number of times the visualisation achieved the top rank, and the column "sc2" shows the same score, excluded the ranks for completion time.

Finally, it must be noted that generalising the results such as those reported by Saket, Endert and Demiralp (2019) to a real-world, decision-making scenario raises many questions. For example, the study design included diagrams with 5 to 34 visual marks, which may not correspond to the demand found in typical decision-making tasks. A visual mark roughly corresponds to a single graphic type (e.g., a single bar in a bar chart). Moreover, there is no guarantee that the performance of a visualisation will be preserved when decision tasks demand the inversion of the structure of an analytic task. To make this point clearer, recall that an analytic task maps the attribute space to the instance space given some concrete conditions on the attributes. However, a decision may require the opposite mapping: given an instance, determine if its attributes satisfy some concrete conditions. Case in point: deciding whether an image represents a cat or a dog (a binary classification problem). Also, there is no guarantee that the performance of a visualisation will be preserved when interactivity is enabled (or disabled). All in all, efforts to quantify the effectiveness of visualisations in decision-making tasks seem to be a fruitful topic in visualisation research.

3.4 Summary and closing remarks

In this chapter, we clarified the gap in the literature on recommender systems considering their application as an assistive technology in gerontological primary care: they should be able to explore data from standardised assessments and provide faithful explanations to an expert-in-the-loop. The goals of such recommenders should align with the consensus mentioned in Section 2.1: (a) guide the selection of interventions to restore or preserve health, (b) predict outcomes, and (c) monitor clinical change over time.

We briefly reviewed the literature on multilabel classification and label ranking tasks because they will be used to formalise the task of creating personalised care plans. In Section 3.2.1, we introduced small contributions to make our exposition clearer. The contributions, which derive from our decision to integrate notation and data encoding, are: (a) the combinatorial characterisation of the multilabel ranking problem in Equation 3.1; (b) a normalised version of the maximum imbalance ratio introduced in Equation 3.2; and (c) the downgrade operator in Equation 3.3, which highlights the "is-a" relationship between these tasks. Finally, we reviewed the literature on visualisation research in search of efficient visualisations in decision-making tasks. We found that the literature does not offer empirical guidance on the selection of visualisations for decision-making tasks, but it does so for analytic tasks. Once again, we introduced a small contribution by connecting these two tasks in order to inform our evaluation of interpretability of visual explanations with empirical evidence. We decided not to approach the topic of interpretability of visual explanations in Sections 3.1 or 3.3 because the explanation style we developed does not seem to match the artefacts that are approached in these literatures.

4 A RECOMMENDATION MODEL FOR PSYCHOMETRIC DATA

In Chapter 1, we stated that the primary objective of our project is the development of a recommendation model for psychometric data: the model should leverage the structure of the data to produce recommendations and explanations. The explanations should be faithful and interpretable by users who are familiar with the psychometric instrument. We now turn to this objective and detail the proposed recommendation model, which is called **Polygrid** from now on. This exposition is organised around the model's learning and predicting pipelines. Each pipeline step is described separately, from two perspectives: algorithmic, which informs how the input data are transformed into the output data, and diagrammatic, which describes how the input and output data are displayed in the diagram. This choice of exposition aims to argue for faithfulness of the explanation diagrams

The remainder of this chapter is organised as follows. First, our usage of the term "psychometric data" is made more precise by presenting the criteria the data must satisfy to be included in this category (Section 4.1). The whoqol dataset is shown to meet the criteria, and a sample which will be used in a running example is drawn and described. In Section 4.2, we describe how Polygrid learns to perform multilabel classification tasks. The learning and predicting pipelines are described, and the impact of alternative choices of hyperparameters is discussed. Section 4.3 is similarly structured, but details how Polygrid learns to perform label ranking tasks. Having covered a number of examples of how Polygrid learns, predicts, and produces explanations, we dedicate Section 4.4 to present a theoretical basis for how Polygrid learns, and Section 4.5 to discuss in which sense the claim that explanations generated by Polygrid are faithful and interpretable is valid.

4.1 Data requirements and data preparation

As seen in Chapter 2, psychometric instruments are designed to measure human attributes. We focus on instruments that follow a factor-analytic approach: the instrument is devised to measure a latent variable, such as quality of life or depression, by means of a questionnaire whose items elicit responses (from members of the target population) that correlate positively with the variable of interest. More precisely, we focus on instruments conceived and developed based on the measurement model in Figure 8. This fact has two relevant implications for data collected with such instrument: the pairwise correlation between domains of the instrument is positive (Eq. 2.2) and the scores obtained for any instrument's domain are quantitative and nonnegative (Eq. 2.6). In practical terms, this means that a dataset of assessments collected with the instrument does not have columns with negative values or constant columns, and variables measured at the nominal level, such as sex, race, religion, or ethnicity, are not used as indicators in the instrument.

Considering a dataset with the characteristics just described, we now detail the additional requirements the data must satisfy. Let $\mathring{D} = (\mathring{X}, Y)$ represent the dataset that holds the original assessments $\mathring{X}$ and their assignments Y. The assessments in $\mathring{X}$ must be organised in a (m, d)-matrix, in which m is the number of individuals, and d is the number of domains of the instrument. Thus, $\mathring{X}_i = (\mathring{x}_{i0}, \ldots, \mathring{x}_{i,d-1})$, with $\mathring{x}_{ik}$ representing the score obtained by the i-th individual for the k-th domain surveyed by the instrument. The assignments in Y are encoded as an (m, n)-matrix, with n being the number of labels in the problem space. For a multilabel assignment, the usual encoding must be followed: $y_{ij} = 1$ indicates that the i-th individual is assigned to the j-th label, and $y_{ij} = 0$ indicates the opposite. For a label ranking assignment, we adopt the same encoding used in benchmark datasets. For instance, $Y_i = (1, 2, 0, 3)$ means that the i-th individual is associated with the ranking $1 \succ 2 \succ 0 \succ 3$. Recall from Section 3.2.1 that the relation $\succ$ represents preference: $a \succ b$ means that a is preferred to b. To encode incomplete rankings, -1 is used as a filler. For instance, $Y_i = (1, 2, -1, -1)$ encodes the incomplete ranking $1 \succ 2$.

The last step to ensure the dataset $\mathring{D} = (\mathring{X}, Y)$ can be handled by Polygrid models consists in scaling the assessments in $\mathring{X}$ to the unit hypercube $[0, 1]^d$. In other words, Polygrid requires a dataset $D = (X, Y)$ such that $x_{ik} \coloneqq \mathring{x}_{ik}/max(\mathring{X}_{:k})$. This requirement secures an important assumption: once differences in range among domain scores are removed, namely $rng(X_{:k}) \subseteq [0, 1]$ for all $k \in 0 \ldots d-1$, all domains can be assumed a priori to be equally relevant to explain the assignments, leaving to the learning model the work of identifying the degree to which each domain can explain the assignments.

Having clarified the requirements a dataset must satisfy so that it can be handled by Polygrid models, we now show that the whoqol dataset fits the bill. Throughout this chapter, this dataset will be used to illustrate how Polygrid learns from data. It contains quality of life assessments of 100 individuals. The assessments were made by researchers affiliated to the Department of Gerontology at the Federal University of São Carlos, Brazil, in October 2019 (Lorenzi *et al.*, 2022). The WHOQOL-BREF instrument was employed. This instrument was detailed in Section 2.4.1. In brief, this was developed by the WHO to assess quality of life across culturally diverse populations (Orley *et al.*, 1996). The instrument has a questionnaire of 26 items, which are answered in 5-points Likert scales, and the majority of items is clustered around four domains: physical health (7 items), psychological (6 items), social relationships (3 items), and environment (8 items). Roughly speaking, the score for each domain is computed as the average of the responses given to items associated with that domain, then multiplied by 4. Thus, domain scores range from 4 and 20 inclusive. The results shown in Table 9 attest that the whoqol dataset satisfies the requirements regarding scores and pairwise correlations (Equations 2.2 and 2.6).

Since this dataset originally did not have any assignments, we created two synthetic assignments to extend it. These are quite simple assignments because they will be used to

Table 9 – Statistics of the whoqol dataset

	Global			Correlations		
Domain	Min.	Avg.	Max.	Psychological	Social	Environment
Physical	10.9	13.9	17.1	0.516	0.326	0.407
Psychological	10.0	14.1	18.0	-	0.548	0.590
Social	5.3	14.9	20.0	-	-	0.408
Environment	7.0	14.9	20.0	-	-	-

Source: The author.

Note: The statistics under the Global heading consider the whole dataset, but the ones under the Correlation heading consider the training set only.

illustrate ideas throughout this chapter. The first is a multiclass assignment based on a stratification method proposed by Silva *et al.* (2014). The method employs the sum of the domain scores of an assessment (aka, the sum-score in Equation 2.4) and a cutoff value to specify a decision boundary. The latter separates the members of a population into two disjoint groups: those with perceived good quality of life (Good QOL, sum-score ≥ 60) and those with perceived poor quality of life (Poor QOL, sum-score < 60). As a result, 41 individuals were assigned to "Good QOL" group, and 59 to the "Poor QOL" group. The second assignment is a label ranking assignment. It was created by applying fuzzy clustering to the assessment scores to obtain four clusters, each identified with a label. Then, each assessment was assigned to a ranking composed of two labels, representing the clusters to which that assessment obtained the highest degrees of membership. Table 10 shows the assessments and corresponding assignments for three individuals in the dataset, which will be used in the running example in the next sections.

Table 10 – Description of the whoqol dataset's instances used in the examples

	$\mathring{X}$				X				Y (ML)		Y (LR)			
i	$\mathring{x}_{i0}$	$\mathring{x}_{i1}$	$\mathring{x}_{i2}$	$\mathring{x}_{i3}$	x_{i0}	x_{i1}	x_{i2}	x_{i3}	y_{i0}	y_{i1}	y_{i0}	y_{i1}	y_{i2}	y_{i3}
89	13.1	11.3	12.0	10.5	0.767	0.630	0.600	0.525	0	1	0	2	-1	-1
98	14.3	14.7	14.7	15.0	0.833	0.815	0.733	0.750	0	1	1	2	-1	-1
30	14.9	15.3	17.3	16.0	0.867	0.852	0.867	0.800	1	0	3	1	-1	-1

Source: The author.

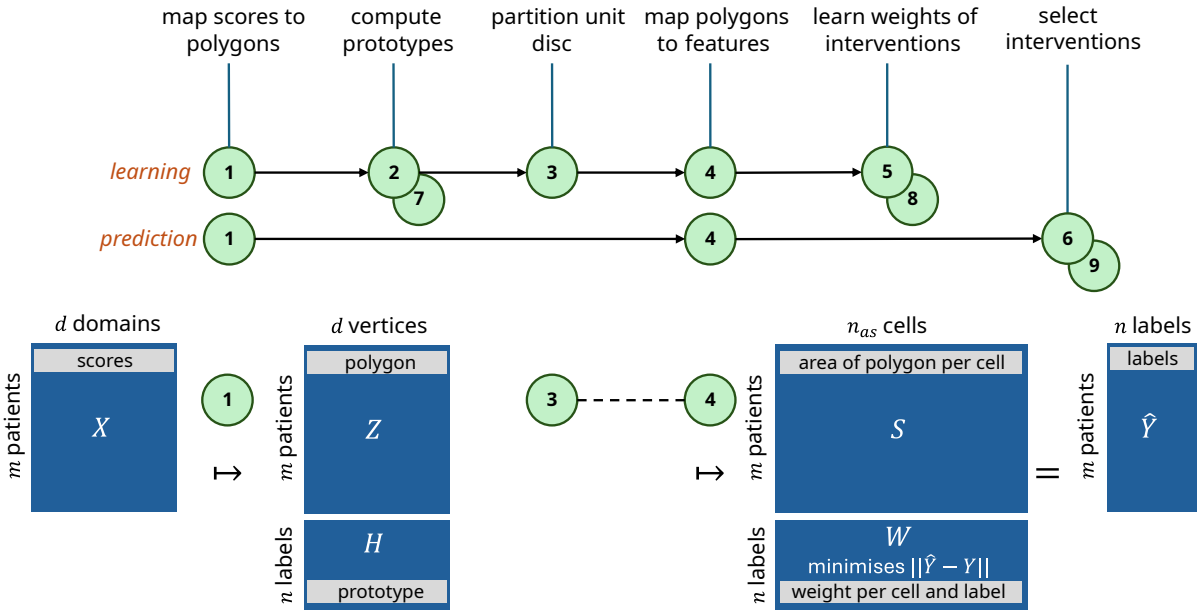
Legend: Column "i" shows the numeric key that was assigned to the individual in the dataset. Columns grouped under "$\mathring{X}$" show the original scores of the assessment, and the ones under "X" show the scaled scores. For both $\mathring{x}_{ik}$, x_{ik}, the order of the domains is the following: physical health ($k=0$), psychological ($k=1$), social relationships ($k=2$), and environment ($k=3$). Columns grouped under "Y (ML)" represent the multilabel assignment: ($y_{i0}=1$) indicates the individual is associated with the "Good QOL" label, and ($y_{i1}=1$) indicates association with the "Poor QOL" label. Finally, columns under "Y (LR)" describe the label ranking assignment: $Y_i=(0,2,-1,-1)$ indicates that the i-th the individual is associated with the incomplete ranking $0 \succ 2$. Each individual is assigned to the two labels that best describe her assessment of quality of life.

4.2 The Polygrid model in multilabel classification tasks

We seek to apply Polygrid to assist a care professional in the creation of a care plan in a gerontological primary care setting. In this niche, the order in which the recommended interventions appear in the care plan may or may not be important. When order matters, the problem can be formalised as a label ranking problem, as seen in Section 3.2.2. When it does not matter, it can be formalised as a multilabel classification problem. In this section, we detail how Polygrid produces recommendations when order is irrelevant.

This section is divided into three parts. The first two are dedicated to detail the learning and prediction pipelines illustrated in Figure 10. During the learning stage, Polygrid goes through five steps to learn the weights that describe the assignment labels. These steps are detailed in Section 4.2.1. To make predictions, Polygrid goes through the three steps detailed in Section 4.2.2. To keep the exposition simple, in both sections, the operation of the model is detailed for a default configuration of hyperparameters. Alternative configurations, as well as their impact on the performance of the model, are contrasted with the default configuration in Section 4.2.3.

Figure 10 – The learning and prediction pipelines of the Polygrid model



Source: The author.

Legend: The learning and prediction pipelines are depicted at the top of the diagram, as two directed paths. Each node corresponds to an algorithm that is described in their corresponding section. The steps representing the learning pipeline for multilabel classification tasks are indicated by the nodes 1 to 5. The learning pipeline for label ranking tasks uses the "shadow" nodes: 1, 7, 3, 4, and 8. A similar reasoning was applied to encode the prediction pipeline. The bottom of the diagram depicts how the input data are transformed along the learning pipeline. For example, the description matrix X is transformed into the matrix of polygons Z by the algorithm corresponding to the step 1.

4.2.1 Learning from data

Roughly speaking, during the learning stage, Polygrid computes the data structures that are needed to render an explanation diagram. These data structures are either related to the assessments in X, to the assignments in Y, or to the feature space. We will refer to Figure 11 to point where these structures are visually depicted on the diagram.

The first step consists in mapping the assessments in X, whose rows are points in the unit hypercube $[0,1]^d$, into polygons in the closed unit disc centred at the origin of the complex plane $\overline{\mathbb{D}}$. This mapping produces simple, solid polygons[1] whose vertices are scalar multiples of the complex roots of unity, $\zeta^d = 1$, confined to the unit disc $\overline{\mathbb{D}}$. In other words, it maps $X_i = (x_{i0}, \ldots, x_{i,d-1})$ to $Z_i = \triangle(x_{i0}\,\zeta_0, \ldots, x_{i,d-1}\,\zeta_{d-1})$, with ζ_k being ordered in the usual fashion[2]. Example: for $d = 4$, there are four roots of unit: $(\zeta_0, \zeta_1, \zeta_2, \zeta_3) = (1, i, -1, -i)$. Thus, $X_a = (x_{a0}, x_{a1}, x_{a2}, x_{a3})$ is mapped to $Z_a = \triangle(x_{a0}, ix_{a1}, -x_{a2}, -ix_{a3})$. Recall that $x_{ik} \geq 0$ by design and $x_{ik} > 0$ by circumstance because the whoqol dataset has only strictly positive scores (see Table 9). These constraints ensure that Z_i is not a degenerate polygon. Since there is no standard name for this family of polygons, and most polygons handled by Polygrid are of this family, we will refer to them as polygons and reserve the term "polygonal shape" to describe a closed polygonal chain (and the area it delimits) in general. Note that the angle between two consecutive roots ζ_k and ζ_{k+1} is constant and equal to $2\pi/d$. This fact is explored in Algorithm 1, which details this step. The diagram in Figure 11 illustrates two assessments that were described in Table 10 for individuals $i = 89$ (labelled as "Poor QOL", $y_{i1} = 1$) and $i = 30$ (labelled as "Good QOL", $y_{i0} = 1$). Thus, the two polygons in the assessment charts in Figure 11 correspond to Z_{89} and Z_{30}.

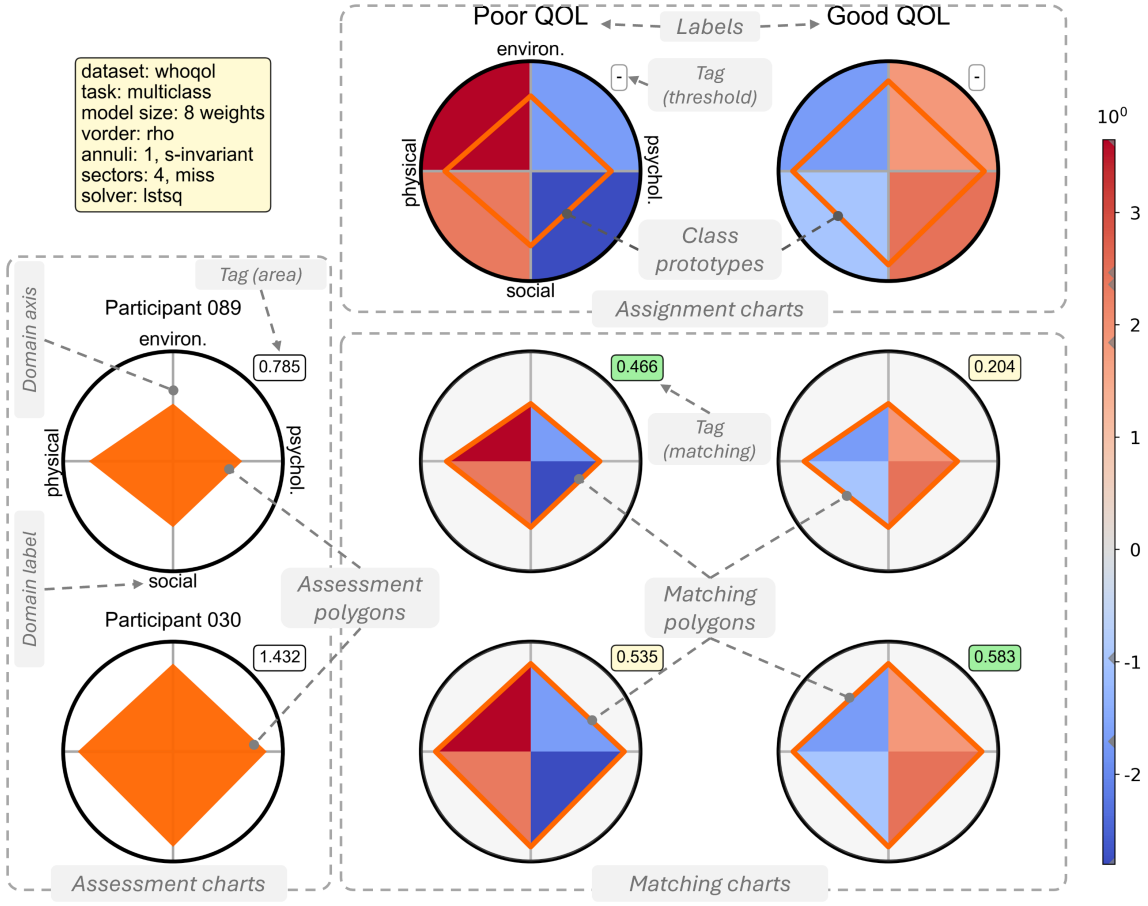
Algorithm 1: Map assessments to polygons (general)

Data: (m, d)-array X
Result: (m, d)-array Z, d-array ζ
Function *uh-to-ud*(X) **is**
  $\theta \leftarrow 2\pi/d$;
  for $k \in 0 \ldots d-1$ **do**
    $\zeta_k \leftarrow \cos(k\theta) + i\sin(k\theta)$; `/* finds d-th roots of unity */`
  end
  for $i \in 0 \ldots m-1$ **do**
    $Z_i \leftarrow \triangle(x_{i0}\,\zeta_0, \ldots, x_{i,d-1}\,\zeta_{d-1})$; `/* polygon on the complex plane */`
  end
  return (Z, ζ) ;
end

[1] A simple polygon is a planar shape whose boundary is a simple polygonal chain (i.e., a chain that is not self-intersecting) that is closed, and a solid polygon is a planar shape that includes both the boundary and the interior of a simple polygon, and its interior does not have holes.

[2] The usual ordering of the roots of unit is anticlockwise, from $\zeta_0 = 1$. The symbol $\triangle$ indicates that the tuple $(z_0, \ldots, z_{d-1})$, taken as the closed chain $(z_0, \ldots, z_{d-1}, z_0)$, forms a solid polygon.

Figure 11 – The anatomy of an explanation diagram generated by the Polygrid model



Source: The author.

Legend: An explanation diagram is composed of a number of radar charts that are disposed on a rectangular grid. The diagram being displayed has been annotated with overlay elements (in grey) to single out its constitutive elements. There are three types of charts: (a) the assessment charts, which are placed on the first column, (b) the assignment charts, on the first row, and (c) the matching charts. The assessment chart depicts the results of a patient assessment. The vertices of an assessment polygon represent the scores the patient obtained for each domain of the instrument. Each assessment chart has a tag, which informs the area of its corresponding polygon. Typically, an explanation diagram displays the assessment of a single individual, but this example includes an extra one to benefit our discussion. The assignment chart depicts the representation of a label in the feature space: the colours in the background correspond to the weights ascribed to different cells of the unit disc. The polygon in an assignment chart, whose vertices are the mean scores of all assessments assigned to its respective label, is called the class prototype. Each assignment chart has a tag, which informs the threshold value for its respective class. In multiclass assignments, this value is omitted because each case is assigned to a single label, as described next. In a matching chart, the polygon is a copy of the assessment polygon on the same row, except that it is filled with "the colours" of the assignment chart in the same column. Each matching chart has a tag, which informs the weighted area of its polygon. A green tag indicates that this value is greater than its respective threshold, and the case is assigned to the respective label. A yellow tag indicates the opposite. In multiclass assignments, the case is assigned to the label with highest weighted area. Finally, the text box at the top left of the diagram describes the default configuration of hyperparameters used to generate the explanation.

The second step consists in computing the class prototype corresponding for each label in the assignments. The class prototype of the j-th label is defined as the mean scores of all assessments that are associated with the label j. Algorithm 2 details the operation. Class prototypes are represented as polygons in the assignment charts of the diagram in Figure 11. Note that the chart corresponding to the label "Poor QOL" ($j=1$) is shown at the left of the one for label "Good QOL" ($j=0$). This inversion occurs because the charts describing assignments, which are placed on the first row of the diagram, are ordered by the size (area) of the polygons that represent the class prototypes, from smallest to largest. This choice of layout is arbitrary, but useful if the labels have an ordinal structure.

Algorithm 2: Compute the class prototypes (multilabel)

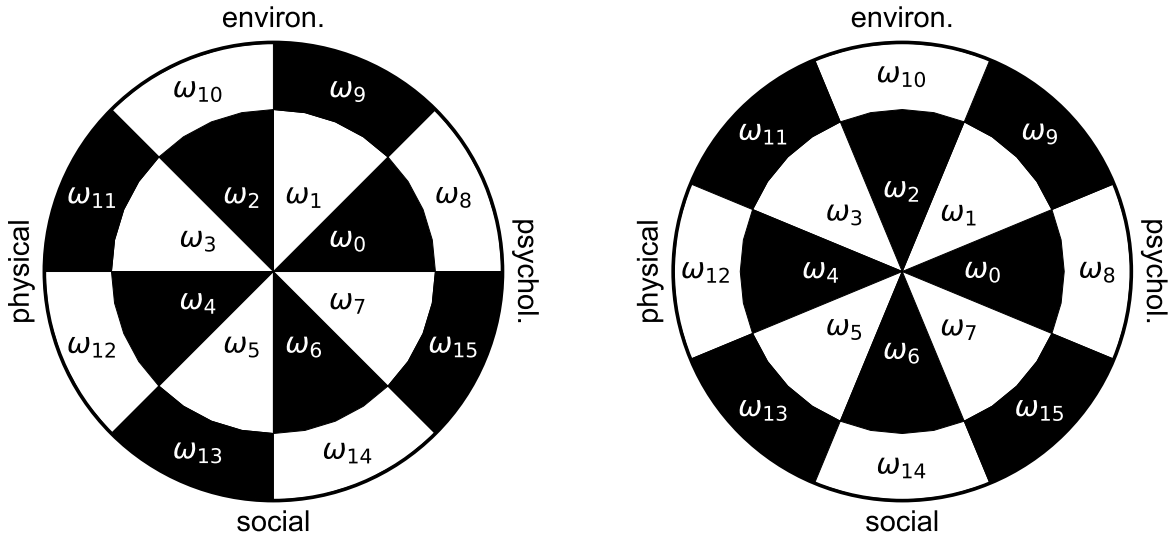
Data: (m, d)-array X, (m, n)-array Y, d-array ζ
Result: (n, d)-array H

1 **Function** *compute-class-prototypes*(X, Y, ζ) **is**
2 **for** $j \in 0 \ldots n-1$ **do**
3 $\mathcal{I} \leftarrow \{i \mid y_{ij} = 1\}$; `/* collects the rows assigned to label j */`
4 $A_j \leftarrow \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} X_i$;
5 $H_j \leftarrow \text{⩓}(a_{j0}\,\zeta_0, \ldots, a_{j,d-1}\,\zeta_{d-1})$;
6 **end**
7 return H ;
8 **end**

The third step defines the feature space. It partitions the unit disc $\overline{\mathbb{D}}$ into disjoint cells that result from the intersection of disjoint annuli and disjoint sectors of the disc. The number of annuli, n_a, and the number of sectors per domain of the instrument, n_s/d, are hyperparameters of the model. Thus, the disc is partitioned into $n_{as} = n_a \times n_s$ annular sectors. These cells are enumerated in anticlockwise order, from the origin to the boundary of the disc. For example, Figure 12a shows a partition of the disc with two annuli and eight sectors. Their pairwise intersection defines $n_{as} = 16$ cells, which are enumerated in the partition $\Omega = (\omega_0, \ldots, \omega_{15})$. This operation is detailed in Algorithm 3. In the pseudocode, the annuli are specified in line 3. Each annulus is described as the difference between two sets: $\{z \in \overline{\mathbb{D}} : |z| \leq \sqrt{(p+1)/n_a}\}$, which describes the set of points contained in the closed disc of radius $\sqrt{(p+1)/n_a}$, excluding the points in $\{z \in \overline{\mathbb{D}} : |z| \leq \sqrt{p/n_a}\}$, the closed disc of radius $\sqrt{p/n_a}$, both centred at the origin. We follow a similar approach to specify the sectors. In line 7, a sector is specified as the set $\{\rho e^{i\theta} : \rho \in (0, 1],\ \theta \in [\,q\,\Delta\theta, (q+1)\,\Delta\theta\,)\}$, which describes all points in the unit disc whose argument $\theta \in [\,q\,\Delta\theta, (q+1)\,\Delta\theta\,)$. Finally, in line 12, the annular sector ω_r is defined as the intersection between an annulus and a sector. This partition defines cells that have the same area and sectors that "miss" the domain axes. The alternative partition shown in Figure 12b has the same number of cells, but its sectors are placed differently, and the cells do not have the same area. These are arbitrary but consequential decisions, and alternatives will be explored in Section 4.2.3.

Figure 12 – The unit disc partitioned into annular sectors

(a) Partition using Algorithm 3 (b) Partition with alternative configuration



Source: The author.

Legend: The unit disc partitioned into the annular sectors defined by $n_a = 2$ annuli and $n_s/d = 2$ sectors per domain. The WHOQOL-BREF instrument has $d = 4$ domains, so the partition has $n_s = 8$ sectors. (a) The widths of the annuli (i.e., the difference between external and internal radii) were chosen so that all cells have the same area. Sectors are placed so that they miss the domain axes. (b) Sectors are placed so that the domain axes are covered. The radius dividing the annuli is $r = 0.5$, so cells have different areas.

Although we use the language of sets to specify regions of the unit disc, which is precise, compact, and convenient for the pseudocode, in practice, we take advantage of open-source libraries for computational geometry, such as Shapely (Gillies *et al.*, 2022). These libraries have geometric primitive operations that allow us to use polygonal shapes to approximate annuli and sectors. This means that, during runtime, each cell ω_r stores the description of an annular sector as a polygonal shape. Of course, this is an approximation, and the goodness of this approximation can be empirically calibrated considering the operations that Polygrid performs on polygonal shapes: difference, intersection, and computing areas. For example, let $\mu(\cdot)$ be a function that "measures" the area of an arbitrary polygonal shape. An implementation of $\mu(\cdot)$ based on the Shoelace theorem will do well in this setup (Lee; Lim, 2017; Needham, 1997/2012, p. 30). Then, a calibration criterion can be $\pi - \sum_r \mu(\omega_r) < \delta$ or $\sum_i \mu(Z_i) - \sum_r \mu(Z_i \cap \omega_r) < \delta$ for a small $\delta > 0$.

In the diagram shown in Figure 11, the partition Ω is not represented directly, but the background colours in the assignment charts reveals their structure: boundaries of annular sectors are wherever colour transitions occur. Although the background colours are not determined by the partitioning, there is a mapping from Ω into the colour palette that represents the weights given to the cells, as we shall see in step 5.

Algorithm 3: Partition the unit disc into annular sectors (general)

Data: integers n_a, n_s
Result: n_{as}-array Ω

```
1  Function partition-ud(n_a, n_s) is
2      for p ∈ 0 ... (n_a − 1) do
3          annulus_p ← { z : |z| ≤ √((p+1)/n_a) } \ { z : |z| ≤ √(p/n_a) }, z ∈ 𝔻̄ ;
4      end
5      Δθ ← 2π/n_s ;
6      for q ∈ 0 ... (n_s − 1) do
7          sector_q ← { ρe^{iθ} : ρ ∈ (0, 1], θ ∈ [ q Δθ, (q+1) Δθ ) } ;
8      end
9      for p ∈ 0 ... (n_a − 1) do
10         for q ∈ 0 ... (n_s − 1) do
11             r ← (p × n_s) + q ;        /* anticlockwise, origin to boundary */
12             ω_r ← annulus_p ∩ sector_q ;
13         end
14     end
15     n_as ← n_a × n_s ;
16     Ω ← (ω_0, ..., ω_{n_as−1}) ;    /* enumeration of cells of the unit disc */
17     return Ω ;
18 end
```

In the fourth step, we turn our attention back to the assessments. Recall that, in step 2, the assessments in X were converted into the polygons in Z. The aim now is to describe how the area of each polygon Z_i is distributed over the disc $\overline{\mathbb{D}}$. This operation is detailed in Algorithm 4. The function $\mu(\cdot)$ that appears in line 4 is the same discussed in the previous step: it computes the area of an arbitrary polygonal shape. Thus, the loop in lines 3-5 describes a procedure in which the intersection between some polygon Z_i and each cell ω_r is obtained. The numeric value of the area of each polygonal shape, $\mu(Z_i \cap \omega_r)$, is computed, and the results are stored in $S_i \coloneqq \Big(\mu(Z_i \cap \omega_0),\, \ldots,\, \mu(Z_i \cap \omega_{n_{as}-1}) \Big)$.

In Figure 11, the matching charts show the decomposition of an assessment polygon Z_i into its intersections with the partition cells. In these charts, the polygon is a copy of the assessment polygon Z_i that appears in the same row, but its interior is filled differently. For instance, there are $n = 2$ matching charts for Z_{89}: one for "Poor QOL", one for "Good QOL". In both charts, the assessment polygon Z_{89} is decomposed into four polygonal shapes $(Z_{89} \cap \omega_r)$, each painted with the same colour in which the annular sector ω_r is drawn in the assignment chart associated with the same label.

The fifth (and final) step focuses on ascribing weights to each cell so that we can reconstruct the assignments in Y based on the assessments in X. This is done by finding an approximate solution to the system of linear equations $SW^{\intercal} = Y$, with W being an (n, n_{as})-matrix. This operation is detailed in Algorithm 5. The weights that reconstruct

Algorithm 4: Map polygons to feature vectors (general)

```
   Data: (m, d)-array Z, n_as-array Ω
   Result: (m, n_as)-array S
1  Function ud-to-fs(Z, Ω) is
2  |  for i ∈ 0 ... (m − 1) do
3  |  |  for r ∈ 0 ... (n_as − 1) do
4  |  |  |  s_ir ← μ(Z_i ∩ ω_r) ;     /* area of polygon Z_i covering cell ω_r */
5  |  |  end
6  |  end
7  |  return S ;
8  end
```

the j-th label of an assignment are computed by the loop in lines 2-4. Each row of the weight matrix W is a solution to the problem of minimising the difference between $Y_{:j}$ and its reconstruction $\hat{Y}_{:j} = SW_j^\intercal$, according to criteria set by the cost function *difference(·)*. In lines 5-6, the learned weights are used to obtain the reconstruction $\hat{Y}$, and its extreme values are used to determine the range of candidate values for a classification threshold. Note that $y_{ij} \in \{0, 1\}$, but this constraint does not apply to $\hat{y}_{ij}$. To make them comparable, a decision rule is usually employed. For example, this rule can be based on a threshold: $\hat{y}_{ij} \leftarrow [\![\hat{y}_{ij} \geq \mathit{threshold}_j]\!]$. In line 7, the classification thresholds are determined as a solution to the problem of maximising the similarity between Y and $\hat{Y}$, according to criteria set by the objective function *similarity(·)*. This approach consists of breaking up a multilabel classification task into multiple binary classification tasks and using a regression model to solve the latter, two strategies discussed in Section 3.2.3.

The specific criteria implemented by the *difference* and *similarity* functions were not detailed, as their behaviour is governed by hyperparameters of the model. For now, we define $\mathit{difference}(Y_{:j}, S, W_j) \coloneqq \|Y_{:j} - SW_j^\intercal\|_2^2$, and highlight that the matrix S is not centred and does not include an extra dimension to encode the intercepts. Regarding the *similarity* function, we define it as $\mathit{similarity}(Y, \hat{Y}, t) \coloneqq F_{1\mu}(Y, [\![\hat{Y} \geq t]\!])$, with $F_{1\mu}$ being the micro-averaged F_1 score described in Section 3.2.3. Once again, these are arbitrary, but consequential decisions and alternative decisions will be explored in Section 4.2.3.

In the diagram in Figure 11, the learned weights are represented as the background colours that appear in the assignment charts. To be more exact, each cell is mapped to a weight by $\psi(j, \omega_r) \coloneqq w_{jr}$, and weights are mapped to colour specs by $\chi(w_{jr})$. For example, the weights given to the cells of the assignment chart for "Good QOL" ($j = 0$) comes from W_0, and the ones for "Poor QOL" ($j = 1$) from W_1. Each assignment "disc" has four cells, so the model has eight weights. Finally, the colour bar on the right of the diagram represents the map $\chi(\cdot)$, which can be used to decode the colours into weights. A more efficient resource is offered in interactive mode: whenever the user clicks on a cell, marks are added to the colour bar to indicate the position of the model weights on the scale.

Algorithm 5: Learn model weights (multilabel)

Data: (m, n_{as})-array S, (m, n)-array Y, int *granularity*, fn *difference*, fn *similarity*
Result: (n, n_{as})-array W, n-array *thresholds*

1 **Function** *learn-weights(*$S, Y,$ *granularity, difference, similarity)* **is**
2 **for** $j \in 0 \ldots (n-1)$ **do**
3 $W_j \leftarrow \arg\min_{W_j} \mathit{difference}(Y_{:j},\, S,\, W_j)$; `/* solves` $SW_j^\intercal = Y_{:j}$ `for` W_j `*/`
4 **end**
5 $\hat{Y} \leftarrow S\,W^\intercal$;
6 $T \leftarrow \mathit{candidates}(\hat{Y}, \mathit{granularity})$;
7 $\mathit{thresholds} \leftarrow \arg\max_{t \in T} \mathit{similarity}(Y, \hat{Y}, t)$; `/* selects best thresholds */`
8 return $(W, \mathit{thresholds})$;
9 **end**

4.2.2 Making predictions and generating explanations

During the prediction stage, Polygrid roughly recapitulates the sequence of steps it performs in the learning pipeline. In fact, the first two steps of the prediction pipeline correspond to the steps 1 and 4 of the learning pipeline, as seen in Figure 10. Given an unseen assessment x, its scores are mapped to a polygon z on the unit disc, and then z is mapped to a vector s in the feature space. The last step of the prediction pipeline combines the learned weights W and thresholds with the vector s to obtain predictions for all labels. This last step is detailed in Algorithm 6. In line 4, a numeric prediction is obtained for each label and stored in $\hat{y}$. The categorical prediction *labels* is derived from $\hat{y}$ in lines 5-7. In the diagram in Figure 11, the numeric values $\hat{y}_j$ are shown in the tags of the matching charts. A green tag indicates that the label should be assigned to the assessment ($\mathit{labels}_j = 1$), and a light-yellow tag indicates otherwise ($\mathit{labels}_j = 0$). Thus, the model recommends that Participant 089 be associated with the "Poor QOL" label, and that Participant 030 be associated with the "Good QOL" label. These are indeed the associations found in the whoqol dataset for the two participants, as seen in Table 10.

It must be clear by now that all data structures required to render an explanation diagram are unaltered byproducts of the learning and predicting processes performed by Polygrid, and all salient data structures handled by the model are displayed in the explanation diagram. Assessment charts display data from Z (learning stage, step 1), assignment charts show data from H and W (learning, steps 2 and 5) according to a layout specified by Ω (learning, step 3), and matching charts combine data from Z and W (learning, steps 1 and 5) according to a layout derived from Ω, with the content and colour of tags determined by $\hat{y}$ and *labels* respectively (from the prediction stage). It is this one-to-one correspondence that supports our argument about the high level of faithfulness of explanation diagrams generated by Polygrid. However, the argument that addresses the interpretability of these explanations will have to wait until Section 4.5.

Algorithm 6: Make predictions (multilabel)

Data: $(1, d)$-array x, n_{as}-array Ω, (n, n_{as})-array W, n-array *thresholds*
Result: $(1, n)$-array $\hat{y}$, $(1, n)$-array *labels*
Function *predict*$(x, \Omega, W,$ *thresholds*$)$ **is**
  $(z, \cdot) \leftarrow$ *uh-to-ud*(x) ; `/* from step 1 */`
  $s \leftarrow$ *ud-to-fs*(z, Ω) ; `/* from step 4 */`
  $\hat{y} \leftarrow sW^{\intercal}$; `/* line 5 from step 5 */`
  for $j \in 0 \ldots (n-1)$ **do**
    $labels_j \leftarrow 1$ if $\hat{y}_j \geq thresholds_j$ else 0 ;
  end
  return $(\hat{y}, labels)$;
end

4.2.3 Exploring alternative values of the main parameters

The previous sections focused on how Polygrid learns from data, makes predictions, and generates explanations using the default values of the model's hyperparameters. For example, in Figure 11, you may have noticed that the order in which the domain labels appear in the charts is different from the order in which the columns are disposed in the dataset, as described in Table 10. The order is "psychological, environment, physical, and social" in the diagram, but "physical, psychological, social, and environment" in the dataset. It turns out that the order in which the instrument's domains are mapped onto the vertices of an assessment polygon may obtain objects with markedly different sizes. For example, take $X_a = (1, \frac{1}{10}, 1, \frac{1}{10})$, $X_b = (1, 1, \frac{1}{10}, \frac{1}{10})$, and $\zeta = (1, i, -1, -i)$. Then $\mu(X_a \odot \zeta) = 0.2 < 0.605 = \mu(X_b \odot \zeta)$[3]. Although X_b is a rearrangement of the scores of X_a, they induce polygons of different sizes. Further evidence that this choice of representation affects the predictive performance of the model is presented in Table 11. The performance obtained with the original ordering, illustrated in Figure 13, differs from that of the default configuration in Figure 11. An inspection of these diagrams shows another sharp difference. In Figure 11, the assignment chart for "Poor QOL" has positive weights in the left semidisc, and negative weights in the right semidisc, while the opposite occurs in the assignment chart for "Good QOL". This pattern is reversed in the assignment charts in Figure 13.

A brute force approach to seek for the best arrangement of the vertices is of limited reach: for an instrument with d domains, there are $(d-1)!/2$ possible arrangements[4]. An alternative is to adopt a heuristic approach. For instance, in Table 11, the configuration for "Figure 11" sorts the vertices according to their pairwise correlations ("$vorder = rho$"). The sorting process works as follows: take the pair of domains with the highest correlation

[3] Here, $x \odot \zeta$ is the elementwise multiplication, and $\mu(x \odot \zeta)$ is an abbreviation of $\mu(\triangle(x \odot \zeta))$.

[4] Rationale: fix $d > 2$. There is $d!$ permutations of $\{0, \ldots, d-1\}$. Half of these are reflections. E.g, $(012, 120, 201)$ and $(210, 021, 102)$ for $d = 3$. Reflections induce same-sized polygons. Each remaining permutation has d rotations. E.g., $012 \xrightarrow{\tau} 120 \xrightarrow{\tau} 201$ for $d = 3$. Rotations induce same-sized polygons. There remains $d!/(2d) = (d-1)!/2$ potential size-changing permutations.

Table 11 – Performance under different model parameters on the whoqol dataset

	Model Parameters						Performance		
Config	ns/d	na	vorder	sector	annulus	solver	accuracy	f1.ma	hamml
Figure 11	1	1	rho	miss	s-invariant	lstsq	0.592	0.580	0.409
Figure 13	1	1	original	miss	s-invariant	lstsq	0.604	0.593	0.396
Figure 14a	1	1	rho	cover	s-invariant	lstsq	0.637	0.621	0.363
Figure 14b	2	1	rho	cover	s-invariant	lstsq	0.642	0.629	0.357
Figure 15a	1	2	rho	miss	s-invariant	lstsq	0.880	0.861	0.120
Figure 15b	1	2	rho	miss	r-invariant	lstsq	0.983	0.982	0.017
Figure 16a	1	1	rho	miss	s-invariant	ridge	0.978	0.976	0.022
Figure 16b	1	1	rho	cover	s-invariant	ridge	0.983	0.982	0.017

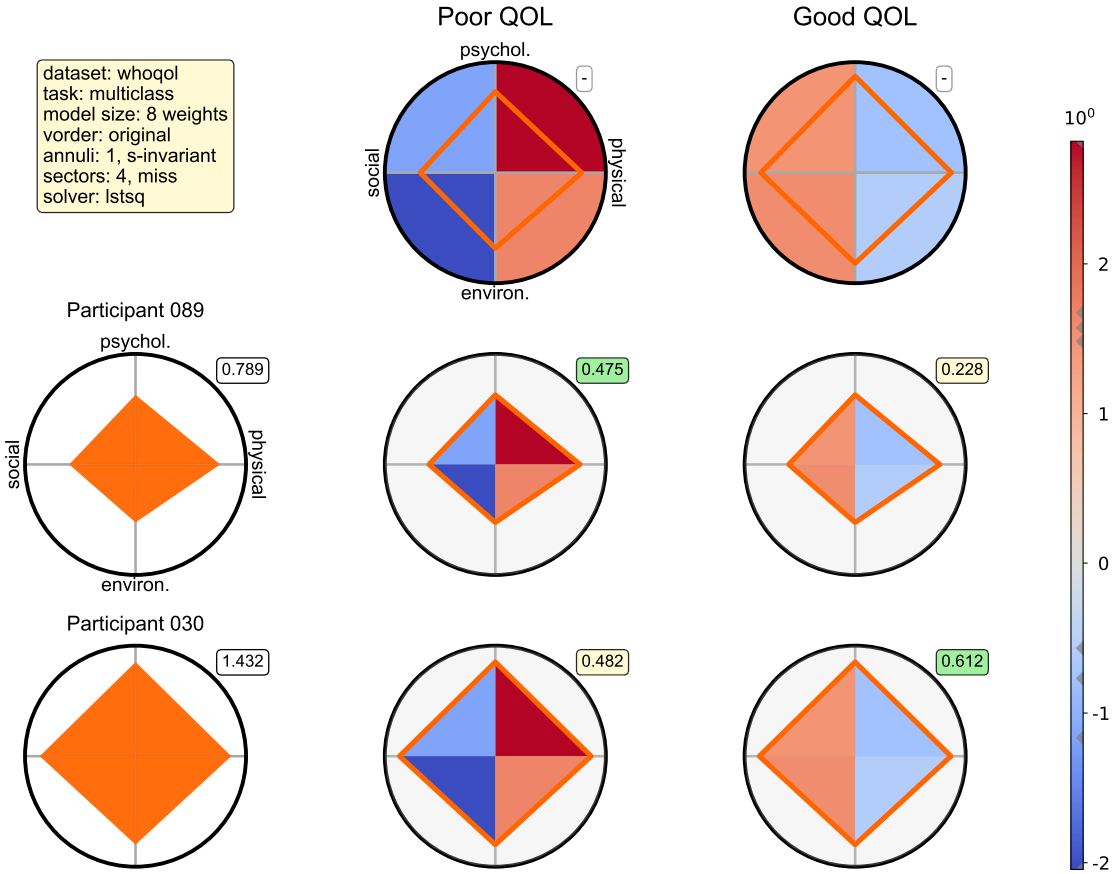
Source: The author.

Legend: Column "Config" indicates the figure displaying an explanation diagram generated by a model operating with the parameters listed under the Model Parameters heading. "ns/d" stands for the number of sectors per domain, "na" stands for the number of annuli, and "vorder" indicates the heuristic used to select the ordering of the domains. Column "annulus" indicates the annulus type: "s-invariant" indicates annular sectors with the same area, and "r-invariant" for annular sectors with the same width. Column "sector" indicates the sector type: "miss" means that the initial sector starts at zero radians, and "cover" means that the first sector is bisected by the first domain axis. Column "solver" indicates the method used to solve systems of linear equations: "lstsq" means least squares, and "ridge" indicates ridge regression. Under the Performance heading, "f1.ma" shows the macro-averaged F1 score, and "hamml" the Hamming loss. The evaluations correspond to the average value obtained from 100 training/test cycles, with identical initial conditions except for differences in parameter values. The parameter values for "Figure 16b" attain the highest performance with less weights.

and name it (a, b). Next, find the domain c that has the highest correlation with one of the endpoints (a or b) and append c to the chain, leading to (c, a, b) or (a, b, c). Repeat the process until all domains have been inserted. We explored other heuristics, but they are all based on the same premises: (a) there is an arrangement that maximises the average size of the assessment polygons, and (b) the larger this average, the higher the discriminability.

Alternative ways of placing sectors over the disc have also been explored. For instance, look at how sectors are placed in the assignment charts in Figure 14a. The figure shows an explanation diagram in which the sectors are placed such that the first domain axis ("psychological") bisects the first sector. This behaviour is determined by the "$sector = cover$" setting shown for Figure 14a in Table 11. This is in contrast with the default configuration used in Figure 11, in which the first sector starts at the first domain axis. In Figure 14a, the sectors are clockwise rotated $\pi/4$ radians. This behaviour is implemented by modifying how sectors are specified in Algorithm 3. More precisely, in line 7, the start of each sector must be clockwise rotated by half of the base angular size, namely $\theta \in [\, q\,\Delta\theta - \frac{\Delta\theta}{2}, (q{+}1)\,\Delta\theta - \frac{\Delta\theta}{2}\,)$. Note that although the colour patterns in the assignment charts are not directly comparable to the ones seen before, they still are opposites to each other, with the extreme positive weight being ascribed to the cell that

Figure 13 – An explanation diagram with the original ordering of the domains



Source: The author.

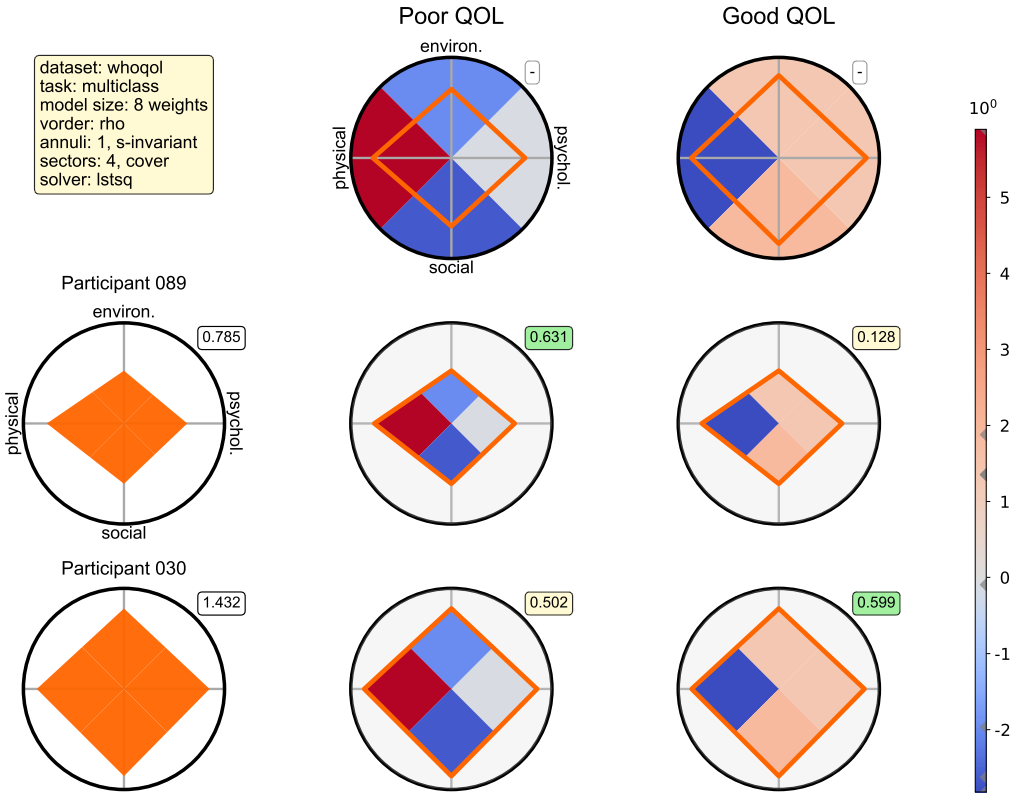
bisects the domain axis labelled "physical" for the "Poor QOL" assignment chart. Also note that rotating the sectors does not change the number of model weights.

Another explanation diagram with rotated sectors is shown in Figure 14b. Its underlying model is identical to the one for Figure 14a, except that it has with twice the number of sectors, and thus twice the number of weights. However, this increase in model size does not translate into a significant increase in performance, as shown in Table 11.

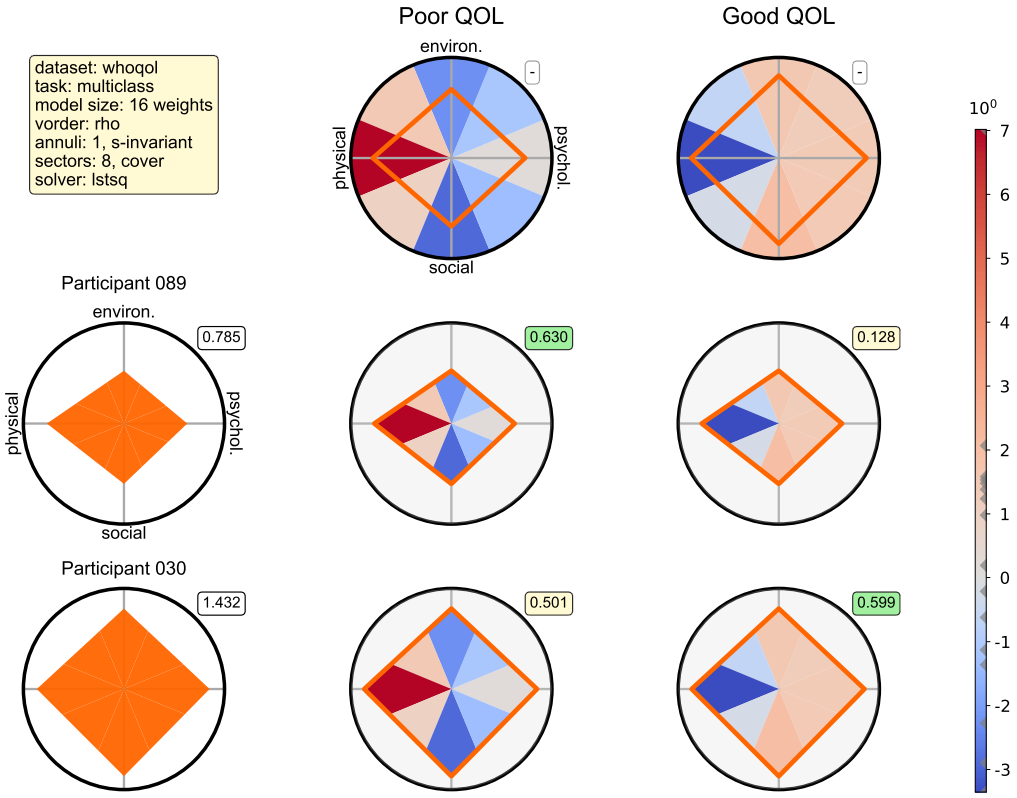
Similarly, alternative ways of specifying how annuli are created have also been explored. In Table 11, it can be seen that the configuration "$annulus = s\text{-}invariant$" is used in Figure 15a, which means that all annuli must have the same area. Since the total area of the unit disc is π, and this area must be equally distributed between two annuli, then each must take up $\pi/2$. In line 3 of Algorithm 3, an annulus is specified as having outer and inner radii of $\rho_{ext} = \sqrt{(p+1)/n_a}$ and $\rho_{int} = \sqrt{p/n_a}$, respectively. This specification implies that, in Figure 15a, the area of the first annulus ($p = 0$) is $\pi(\rho_{ext}^2 - \rho_{int}^2) = \pi(1/2 - 0) = \pi/2$, and the area of the second annulus ($p = 1$) is $\pi(1 - 1/2) = \pi/2$. In general, this specification ensures that the area of any annulus is the constant $\pi(\rho_{ext}^2 - \rho_{int}^2) = \pi/n_a$.

Figure 14 – Examples of an explanation diagram with different types of sectors

(a) using $ns/d = 1$, $sector = cover$



(b) using $ns/d = 2$, $sector = cover$



Source: The author.

An alternative to having annuli with the same area is shown in Figure 15b, in which all annuli have the same width. This behaviour is implemented by modifying the specification in line 3 of Algorithm 3. To be more accurate, the line should now read $annulus_p \leftarrow \{\, z : |z| \leq (p+1)/n_a \,\} \setminus \{\, z : |z| \leq p/n_a \,\}$, $z \in \overline{\mathbb{D}}$. In Table 11, it can be seen that the configuration used in Figure 15b is called "*annulus = r-invariant*".

Note that, differently from the changes in configuration shown up to now, increasing the number and changing the type of annuli translated into significant improvements in performance when compared to the default configuration. This makes sense in light of the decision boundary that was used to create the synthetic assignments. Recall from Section 4.1 that the assignment rule for the multiclass classification task was described as a map from an arbitrary assessment X_i to the "Good QOL" label if its sum-score ≥ 60, and to the "Poor QOL" label otherwise. In the whoqol dataset, given two arbitrary assessments X_a and X_b, if the sum-score for X_a ($\sum_k x_{ak}$) is greater than that of X_b ($\sum_k x_{bk}$), then the area of the polygon representing X_a is greater than the area of the polygon representing X_b for about 99.3% of the pairs[5] The increase in the number of annuli allowed both models to detect this boundary more precisely. In Figure 15a, the class prototype for the "Poor QOL" label is almost inscribed in the first annulus, and in Figure 15b, it is practically circumscribed by the first annulus. This allowed both models to ascribe negative weights to the second annulus, preventing polygons that cover substantial chunks of that area to be classified as a "Poor QOL" case. Note that models in Figure 15 have 16 weights each.

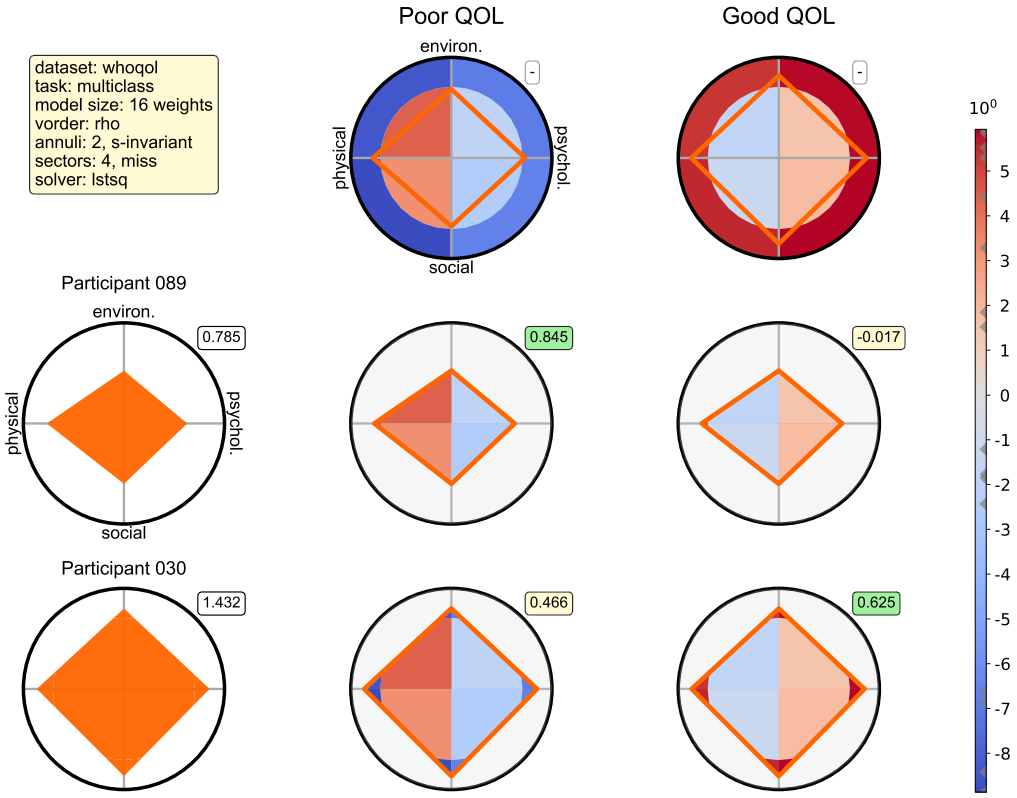
Finally, we move on to describe alternative ways of computing the weights that are ascribed to the partition cells. As stated in Section 4.2.1, the system of linear equations $SW^\intercal = Y$ that is solved during the learning stage (step 5) has two important parameters: the *difference* and the *similarity* functions. In the default configuration, the cost function *difference* is defined as $\mathit{difference}(Y_{:j}, S, W_j) \coloneqq \|Y_{:j} - SW_j^\intercal\|_2^2$. Thus, it does not perform regularisation. Besides, recall that the matrix S, whose rows hold the decomposition of the assessment polygons over the partition of the disc, is a nonnegative matrix. Each element s_{ir} corresponds to the area of the polygon Z_i that is covered by the cell ω_r. This means that the matrix S is not mean-centred, and it does not include an extra dimension to encode intercepts. Of course, these characteristics of the matrix S have important consequences for how the weights must be interpreted. Although mean-centering would imply having $s_{ir} < 0$ for some i and r, which raises issues to the interpretability of the diagram, introducing intercepts or adding a regularisation term are unproblematic changes.

For instance, the explanation diagram in Figure 16a was generated by a model that includes intercepts. As shown in Table 11, the configuration for Figure 16a is identical to the default configuration used in Figure 11, except for the value of the parameter that

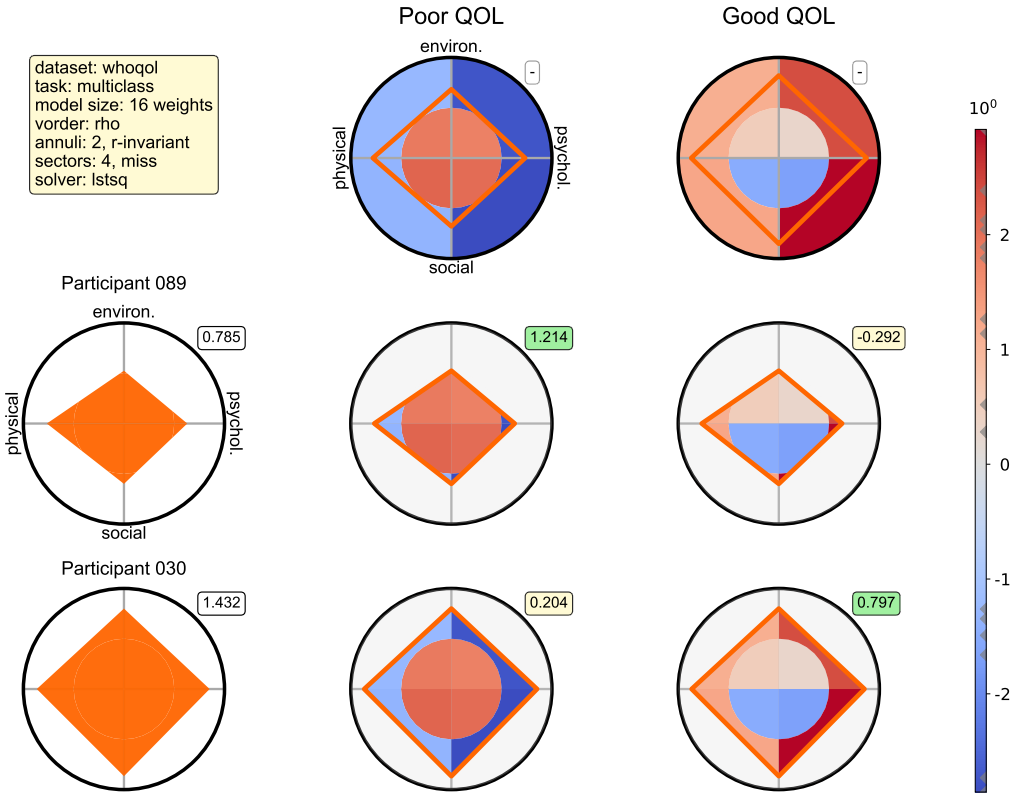
[5] This relationship between sum-scores and the area of its corresponding polygon (or area-score) is developed in Appendix A, which also presents evidence supporting the relationship.

Figure 15 – Examples of an explanation diagram with different types of annuli

(a) $na = 2$, $annulus = s\text{-}invariant$ (all annuli have the same area)



(b) $na = 2$, $annulus = r\text{-}invariant$ (all annuli have the same width)



Source: The author.

indicates that the "ridge" solver must be used. Accordingly, since there are $n = 2$ labels in the assignments, then two additional weights must be introduced. These weights are shown in the grey tags that are placed just below the matching charts (in the last row). This model also introduces regularisation: $\mathit{difference}(Y_{:j},\, S,\, W_j) \coloneqq \|Y_{:j} - SW_j^\intercal\|_2^2 + \alpha\|W_j\|_2^2$.

Regarding the *similarity* function, the main characteristic that may be changed to adapt the model to a given context is its "metric". In the default configuration, this function was defined as $\mathit{similarity}(Y, \hat{Y}, t) \coloneqq F_{1\mu}(Y, [\![\hat{Y} > t]\!])$, with $F_{1\mu}$ being the micro-averaged F_1 score. There may be cases in which replacing this metric with the macro-averaged F_1 score (or other metric) would improve the overall performance of the model.

The explanation diagram in Figure 16b was also generated by a model that includes intercepts. As shown in Table 11, its configuration is identical to the default configuration used in Figure 11, except for the value set for the solver parameter ("*solver=ridge*") and the one set for selection of sector type ("*sector=cover*"). It attains the same performance that was observed for the model that generated the diagram displayed in Figure 15b, which was the best among the reported models, but it needs less weights (10 vs. 16 weights). However, it must be said that these diagrams offer quite different explanations regarding why each assessment should be assigned to their designated labels, and both researchers and practitioners would be justified in preferring one model over another depending on the particular context of their application.

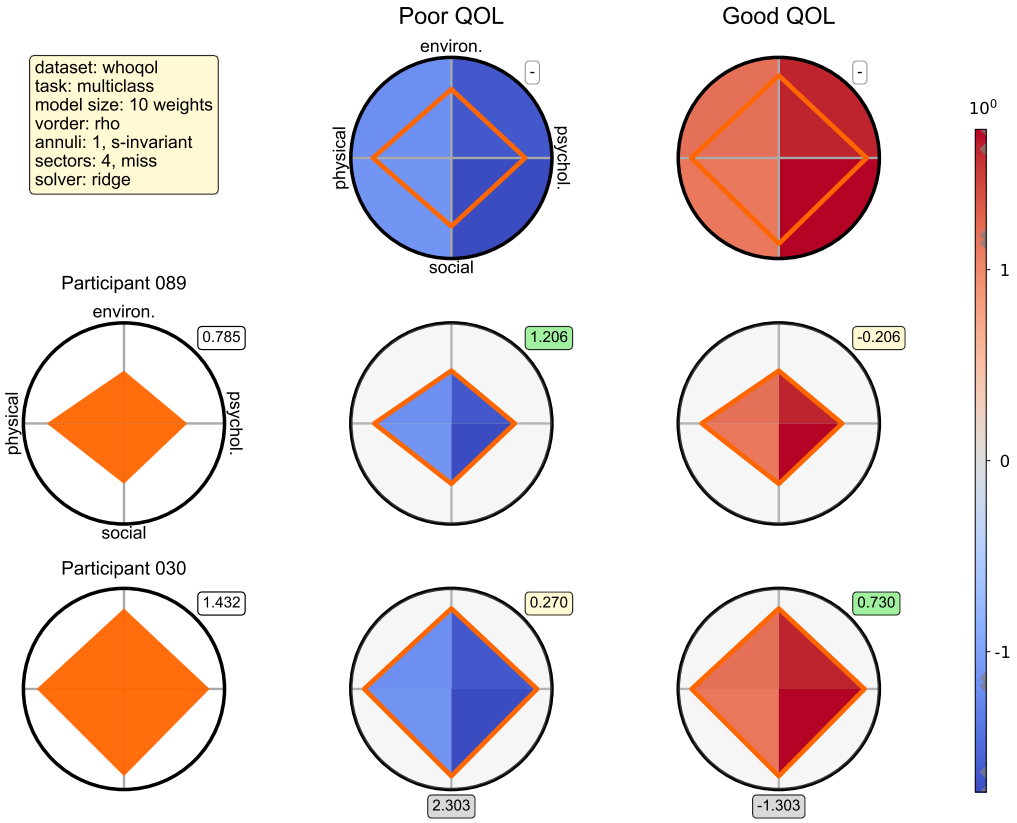
4.3 The Polygrid model in label ranking tasks

To learn how to perform label ranking tasks, Polygrid goes through the same steps that were described for multilabel classification tasks, except that some of the algorithms must be adapted. This is needed because the encoding used to describe label ranking assignments is different from the one used for multilabel assignments. Recall from Section 4.1 that we adopted the encoding that is used in benchmark datasets for label ranking tasks, in which $y_{i\ell}$ encodes which label has rank ℓ according to the preference set to the i-th assessment. For instance, $Y_i = (1, 2, -1, -1)$ encodes the incomplete ranking $1 \succ 2$.

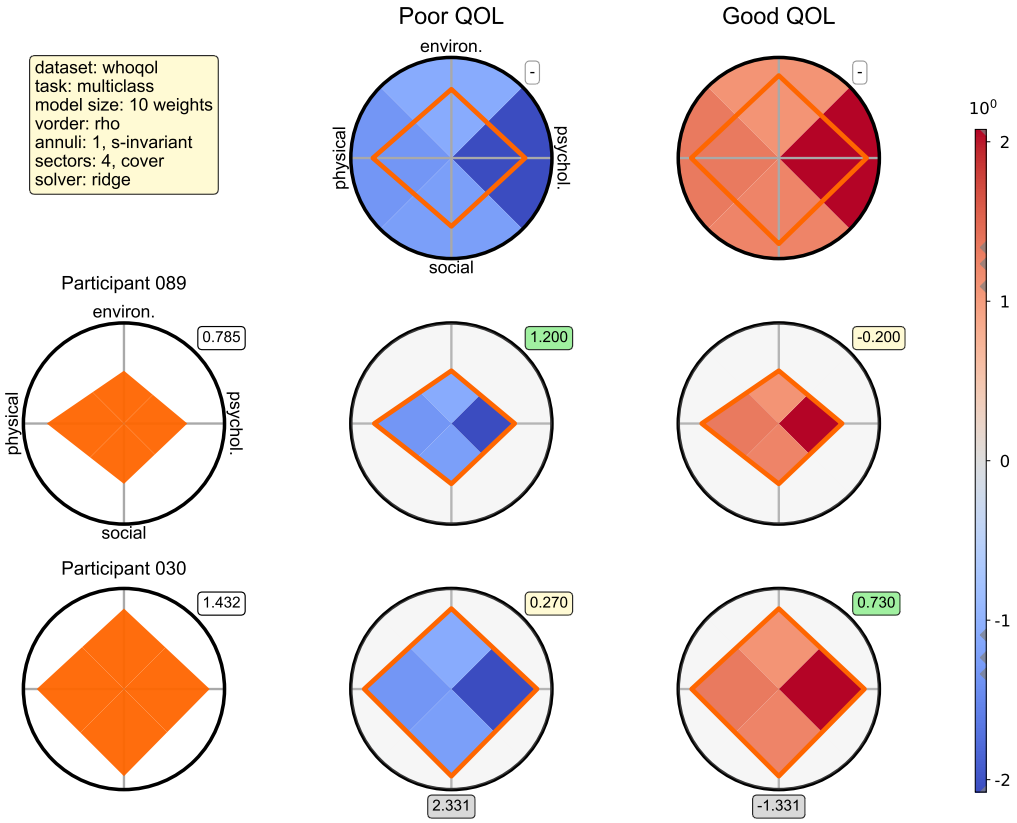
Since we are using the whoqol dataset as a running example, and this dataset has no assignments originally, we created an assignment matrix Y with four labels, and ensured that each assessment is assigned to two labels. This was done by applying fuzzy c-Means clustering (Ross, 2010) to the assessment data X to produce four clusters. The weighting exponent was chosen according to a method by Yu, Cheng and Huang (2004) to prevent instability across repetitions. The clustering results are encoded as a nonnegative (m, n)-matrix U, in which u_{ij} indicates the degree of membership of the i-th assessment to the cluster representing the j-th label, with $\sum_j u_{ij} = 1$. We assigned each assessment X_i the two labels corresponding to the clusters where X_i had the highest degrees of membership. Figure 17 illustrates how assessments in the whoqol dataset cluster around each label.

Figure 16 – Examples of an explanation diagram with different types of solvers

(a) using $sector = miss$, $solver = ridge$

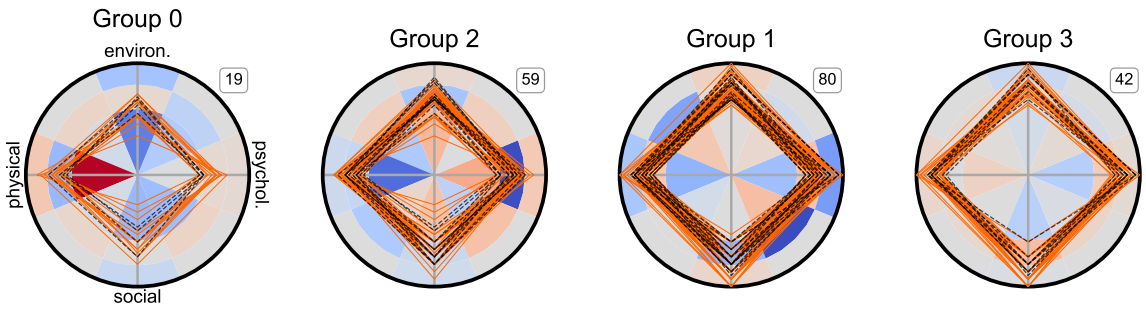


(b) using $sector = cover$, $solver = ridge$



Source: The author.

Figure 17 – Instances of the whoqol dataset clustered around four labels



Source: The author.

Legend: The clustering process generates four labels, from "Group 0" to "Group 3". In the diagram, all assessments that are assigned to a given label have their corresponding polygons superimposed on the radar chart under that label. Training samples are represented as red polygons with solid lines, and test samples are shown as black polygons with dashed lines. The tag at the top right of each radar chart displays the number of assessments assigned to the corresponding label. The whoqol dataset has 100 assessments, and each is assigned to two labels. Thus, the total number of assignments is twice that number.

4.3.1 Learning from data

The learning pipeline for label ranking tasks shown in Figure 10 indicates that Polygrid performs the same steps described for multilabel classification tasks, except for Algorithms 2 and 5, which must be replaced with Algorithms 7 and 8. The adaptation needed to compute class prototypes is detailed in Algorithm 7, line 2: the label ranking assignment Y is converted into a multilabel assignment using the downgrade operator from Equation 3.3. For example, $Y_i = (0, 2, -1, -1)$ is converted into $Y_i' = (1, 0, 1, 0)$. The remainder of the algorithm matches Algorithm 2 line by line. Thus, the class prototype for the j-th label represents the average scores of all the assessments assigned to that label. However, it must be noted that as the density of the assignment Y approaches its upper bound, the class prototypes converge to the same polygon. This has negative effects on interpretability, as the user loses the ability to visually match the assessment polygon and the class prototypes. In that case, alternative adaptation methods should be considered.

The adaptation needed to learn weights and thresholds for label ranking tasks is detailed in Algorithm 8. In line 2, the label ranking assignment Y is transformed into a membership matrix U. The *logranks*$(\cdot)$ function uses the position of a label to define an arbitrary degree of membership that is locally consistent. For that purpose, let's define:

$$u(j, Y_i) := \begin{cases} \dfrac{2^{n-1-\pi(j,Y_i)}}{2^n - 1} & \text{if } j \in Y_i, \\ 0 & \text{otherwise,} \end{cases} \tag{4.1}$$

where $\pi(j, Y_i)$ is the position of the label j in the ranking given by Y_i. For example, let $Y_i = (3, 2, -1, -1)$. The *logranks*$(\cdot)$ function assigns $u_{ij} \leftarrow u(j, Y_i)$ for $j \in 0 \ldots (n-1)$, and

Algorithm 7: Compute the class prototypes (label ranking)

Data: (m, d)-array X, (m, n)-array Y, $(1, d)$-array ζ
Result: (n, d)-array H

1 **Function** *compute-class-prototypes*(X, Y, ζ) **is**
2 $\quad Y' \leftarrow Y^{\downarrow}$; `/* downgrading, see Eq. 3.3 */`
3 $\quad$ **for** $j \in 0 \dots n-1$ **do**
4 $\quad\quad \mathcal{I} \leftarrow \{i \mid y'_{ij} = 1\}$; `/* collects the rows assigned to label` j `*/`
5 $\quad\quad A_j \leftarrow \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} X_i$;
6 $\quad\quad H_j \leftarrow \blacktriangle(a_{j0}\,\zeta_0, \dots, a_{j,d-1}\,\zeta_{d-1})$;
7 $\quad$ **end**
8 $\quad$ return H ;
9 **end**

then normalises the scores so that $\sum_j U_{ij} = 1$. As a result, we end up with $U_i = (0, 0, \frac{1}{3}, \frac{2}{3})$.

In lines 3-5, the algorithm computes the weights as an approximate solution to the system of linear equations $SW^{\intercal} = U$, and the thresholds are computed in lines 6-8. Unlike its counterpart for multilabel classification, the *similarity*$(\cdot)$ function now uses the matrix $\hat{U}$ to produce an initial prediction $\hat{Y}$ in ranking representation. To address the need to handle incomplete rankings, any labels $\hat{y}_{i\ell}$ such that $\hat{u}_{ij} < t$, with ℓ being the position of the label j given by $\ell := \pi(j, Y_i)$, are removed. For instance, let $\hat{U}_i = (-0.05, 0.19, 0.21, 0.37)$. Then, an initial guess for the predicted ranking is $\hat{Y}_i = (3, 2, 1, 0)$. Now take $t = 0.2$. Then, all labels with a membership degree lower than t are removed, resulting in $\hat{Y}_i = (3, 2, -1, -1)$. As in its multilabel counterpart, this is done iteratively and the threshold values that maximise the similarity between Y and $\hat{Y}$ are selected. We used the mean squared error in these comparisons, although alternative metrics could also be considered.

Algorithm 8: Learn model weights (label ranking)

Data: (m, n_{as})-array S, (m, n)-array Y, int *granularity*, fn *difference*, fn *similarity*
Result: (n, n_{as})-array W, n-array *thresholds*

1 **Function** *learn-weights(*S, Y, *granularity, difference, similarity)* **is**
2 $\quad U \leftarrow \mathit{logranks}(Y)$;
3 $\quad$ **for** $j \in 0 \dots (n-1)$ **do**
4 $\quad\quad W_j \leftarrow \arg\min_{W_j} \mathit{difference}(U_{:j},\, S,\, W_j)$; `/* solves` $SW_j^{\intercal} = U_{:j}$ `for` W_j `*/`
5 $\quad$ **end**
6 $\quad \hat{U} \leftarrow S\,W^{\intercal}$;
7 $\quad T \leftarrow \mathit{candidates}(\hat{U}, \mathit{granularity})$;
8 $\quad \mathit{thresholds} \leftarrow \arg\max_{t \in T} \mathit{similarity}(Y, \hat{U}, t)$; `/* selects best thresholds */`
9 $\quad$ return $(W, \mathit{thresholds})$;
10 **end**

4.3.2 Making predictions and generating explanations

The prediction stage must also be adapted to perform label ranking tasks. The adaptation is detailed in Algorithm 9. The first change appears in line 4, in which the feature vector s is combined with the learned weights to obtain the vector $\hat{u}$. This means that $\hat{u}_j$ holds the predicted degree of membership of the assessment x to the cluster underlying the j-th label. The loop in lines 5-7 determines which labels should be included ($label_j = 1$) or removed ($label_j = 0$) from the predicted ranking. This is decided by comparing the degree of membership $\hat{u}_j$ with the learned threshold value. Finally, in line 8, the function $membership2rank(\cdot)$ combines the predicted membership vector $\hat{u}$ with the mask vector *labels* to produce the final prediction $\hat{y}$, encoded as a label ranking.

Algorithm 9: Make predictions (label ranking)

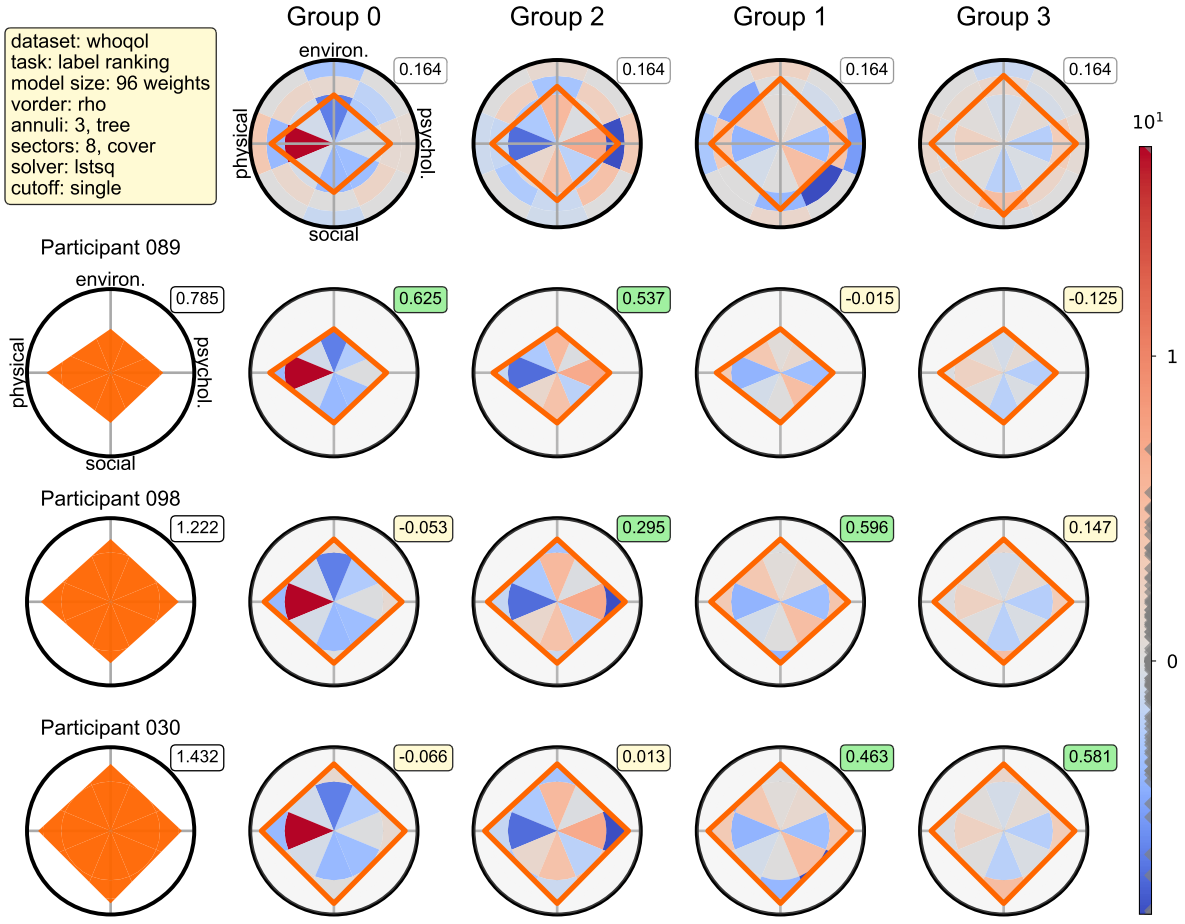
Data: $(1, d)$-array x, n_{as}-array Ω, (n, n_{as})-array W, n-array *thresholds*
Result: $(1, n)$-array $\hat{y}$, $(1, n)$-array *labels*

```
1  Function predict(x, Ω, W, thresholds) is
2  |  (z, ·) ← uh-to-ud(x) ;                          /* from step 1 */
3  |  s ← ud-to-fs(z, Ω) ;                            /* from step 4 */
4  |  û ← sW^T ;                                  /* line 5 from step 5 */
5  |  for j ∈ 0 ... (n − 1) do
6  |  |  labels_j ← 1 if û_j ≥ thresholds_j else 0 ;
7  |  end
8  |  ŷ ← membership2rank(û, labels) ;
9  |  return (û, ŷ) ;
10 end
```

An explanation diagram generated by the Polygrid model trained on the whoqol dataset with label ranking assignments is displayed in Figure 18. The three assessments described in Table 10 are displayed. They are sorted by increasing size of the assessment polygon, from top to bottom. In other words, they are sorted by increasing levels of quality of life. The assignment charts are organised in the same way, from left to right. Similar to the interpretation given to the diagram for multilabel classification, a matching chart with a green tag indicates that the corresponding assessment (the one on the same row) should be assigned to the label (in the same column), and a light-yellow tag indicates the opposite. The value shown in the tag corresponds to the estimated degree of membership $\hat{u}_{ij}$. These values determine the recommended ranking of the labels for each assessment. For instance, the ranking for "Participant 89" is $0 \succ 2$, "Part. 98" has $1 \succ 2$, and "Part. 30" has $3 \succ 1$. Note that the pattern formed by the green tags in the diagram is coherent with the choices made about the ordering of the assessment and assignment charts in the diagram: the individual with the lowest score on quality of life is assigned to the two labels of smaller sizes; the individual with the highest score is assigned to the two labels of larger sizes, and the individual with an intermediary score is assigned the intermediate sized labels.

Finally, the yellow tag on the top left of the diagram lists the Polygrid configuration used to produce the explanation diagram. These configurations are similar to those explored previously, except for the setting for annulus type, which is set to "$annulus = tree$". In this mode, the levels at which the annuli are placed are determined by the thresholds produced by a CART decision tree trained on the data that was submitted to the Polygrid instance (i.e., the training partition of the dataset $D = (X, Y)$). The depth of the tree is constrained to produce as many thresholds as is necessary to create n_a annuli. For example, if the root node of the decision tree rules that an assessment with a score on the psychological dimension less than 0.81 should follow the left branch, then there will be an annulus with the interior/exterior radius set at that threshold. New thresholds are added by visiting the tree in depth-first order, stopping when enough annuli have been specified.

Figure 18 – An explanation diagram produced by Polygrid for a label ranking task



Source: The author.

Legend: An explanation diagram for the three assessments detailed in Table 10. Assignments were created by applying fuzzy clustering on data from the whoqol dataset. The configuration used to tune Polygrid is similar to the ones seen previously, except for the annulus type, which is set to "tree". In this case, the widths of the annuli are selected according to the thresholds found by a decision tree induced from the whoqol dataset. In a sense, this setting may be described as an ordinal multilabel classification of the assessments.

Up to this point in our exposition, we have focused on how Polygrid performs multilabel classification and label ranking tasks. The problem of recommending referrals is equivalent to learning to solve these tasks, as seen in Section 3.2.2. We framed the operation of the model as a processing pipeline through which the data is progressively transformed to extract numerical representations that can be used later to predict assignments from unseen assessments in the dataset. Along the way, the algorithmic description offered for each step of the pipeline was systematically complemented by a description of how these numerical representations are graphically displayed on the explanation diagram. Although this effort demonstrates that explanations generated by Polygrid faithfully reflect the operation of the model, it gives little insight about why the model learns, and it does little to justify any claims about its interpretability. We now turn our attention to these issues.

4.4 The learnability of the Polygrid model

Minute details about the operation of the Polygrid model have been described, and several examples have illustrated its application. The results shown are in agreement with the ground truth in Table 10, which was used in the running example. Although the tasks encoded in the dataset are too simple (as they should be to benefit this exposition), the results convey some evidence in support of the idea that the Polygrid model can learn. Of course, stronger empirical evidence will be provided later, in Chapter 5. The algorithms describing the learning steps give us some clues about how Polygrid works, but we are missing a clearer, deeper, and more concise explanation of why it learns. We now tackle this shortcoming by proposing a theoretical basis to explain why Polygrid should be expected to perform learning tasks such as multilabel classification and label ranking.

The Polygrid model loosely resembles a kernel method (Hofmann; Schölkopf; Smola, 2008). Similarly to the latter, Polygrid maps a d-dimensional vector X_i, which in our context holds the standardised assessment scores obtained by the i-th individual, to a vector $S_i := \Phi(X_i)$, which usually sits in a higher-dimensional space $\mathbb{R}^{n_{as}}$, called feature space in the literature on kernel methods. In Polygrid, this mapping corresponds to:

$$\begin{aligned} \Phi : (0,1]^d &\to [0, \frac{1}{2}]^{n_{as}} \\ X_i &\mapsto \Phi(X_i) := (\textit{ud-to-fs} \circ \textit{uh-to-ud})(X_i), \end{aligned} \tag{4.2}$$

as described in Algorithms 1 and 4. In other words, a vector X_i in the unit hypercube is mapped to a polygon on the unit disc, and then to a point in the Polygrid's feature space.

The intuition here is that, although real-world data most often encode nonlinear relationships $f(X) = Y$, $\Phi(\cdot)$ may be able to extract features from X such that a linear relationship $\Phi(X)W^\intercal = Y$ becomes a reasonable hypothesis. However, the resemblance between Polygrid and traditional kernel methods ends there. In kernel methods, the explicit evaluation of $\Phi(X)$ is bypassed by the *kernel trick*. In this operation, the inner product in

the feature space is replaced by the evaluation of a kernel k in the original domain:

$$\begin{aligned} k : \mathbb{R}^d \times \mathbb{R}^d &\to \mathbb{R} \\ (X_a, X_b) &\mapsto k(X_a, X_b) = \langle\, \Phi(X_a), \Phi(X_b) \,\rangle. \end{aligned} \tag{4.3}$$

This trick is then used by some learning model to solve a task of interest, by handling the Gram matrix of the kernel k (Schölkopf; Tsuda; Vert, 2004; Carrington, 2018; Otto, 2023). In contrast, Polygrid explicitly handles vectors in the feature space, without resorting to a kernel. The dimension of this space is determined by the hyperparameters n_a and n_s, which specify the partitioning of the unit disc described in Algorithm 3. Moreover, Polygrid explores a recurring idea from factorisation methods in recommender systems: it learns an explicit representation in the feature space for each label j, namely W_j, such that $\langle\, S_i, W_j \,\rangle$ approximates Y_{ij} (see Algorithms 5 and 8). In the jargon of recommender systems, the vectors X_i and S_i describe a user, W_j represents an item, and Y_{ij} and the scalar $\langle\, S_i, W_j \,\rangle$ correspond to the relevance of the j-th item to the i-th user. All in all, the degree to which Polygrid can solve learning tasks depends on the mapping $\Phi(\cdot)$ being able to extract features from X that makes $\Phi(X)W^\intercal = \hat{Y}$ a useful approximation of Y.

Ideally, at this point, a proof demonstrating that Polygrid can learn any arbitrary hypothesis $h(X) = Y$ should be presented to the reader. Such a proof would probably require a closed form of $\Phi(\cdot)$ that is amenable to functional analysis, so that an approach similar to those in Cybenko (1989) or Steinwart (2002) could be devised to show that Polygrid can learn a hypothesis h that is arbitrarily close to the ground truth f if n_{as} is sufficiently large. Since we cannot provide such a result, we instead share an insight into the effect of increasing the dimensionality of the feature space on the model's accuracy.

First, note that the relationship between subjects and interventions described by the linear hypothesis $\Phi(X)W^\intercal = SW^\intercal = \hat{Y} \approx Y$ allows us to interpret the inner product $\hat{Y}_{ij} = \langle\, S_i, W_j \,\rangle$ as the degree to which the health assessment of the i-th individual matches the (latent) health profile described by W_j. This interpretation of $\langle\, S_i, W_j \,\rangle$ and of W_j is valid inasmuch as $\hat{Y}$ is a useful approximation to Y. Second, note that the orthogonal projection of S_i on W_j, namely $proj_{W_j}(S_i)$, which corresponds to the additive component of S_i that is parallel to W_j, gives us an insightful interpretation of the vector $\langle\, S_i, W_j \,\rangle W_j$:

$$proj_{W_j}(S_i) = \frac{\langle\, S_i, W_j \,\rangle}{\langle\, W_j, W_j \,\rangle} W_j \implies \langle\, S_i, W_j \,\rangle W_j = \|W_j\|^2\, proj_{W_j}(S_i). \tag{4.4}$$

Of course, $\langle\, S_i, W_j \,\rangle W_j$ maps some point that sits on the linear subspace spanned by W_j, and Equation 4.4 clarifies that this point corresponds to the additive component of S_i that is parallel to W_j, scaled by the squared length of W_j. Considering the interpretation of $\langle\, S_i, W_j \,\rangle$ and W_j as concepts in the application domain, and the results just presented, we developed a graph in which W_j figures as a "ruler" that measures the degree to which a subject matches the health profile that experts associate with the j-th intervention.

This graph is shown in Figure 19, which illustrates the model's performance on the whoqol dataset. As said before, this dataset has $m=100$ assessments and $n=2$ labels. The upper subplot depicts all matching scores for the "Good QOL" label, namely $\langle\, S_i, W_0 \,\rangle$ for all $i \in 0 \ldots (m-1)$, and the lower subplot does the same for the "Poor QOL" label, by depicting the matching scores $\langle\, S_i, W_1 \,\rangle$. Thus, the graph depicts mn points in total[6].

The thick line that runs through each subplot depicts a compact set within the subspace spanned by its respective vector W_j. Since this subspace is "a straight line" no matter the dimension of the space in which it is embedded, the depiction of a compact subset of this space as a line segment in the "paper" plane is warranted by an affine mapping. The graduation on the abscissa identifies the specific set being depicted.

In the upper half of the "Good QOL" subplot, points are organised in two levels, which are identified in the ordinate axis as "tr/TN" and "tr/TP". The prefix "tr" indicates that these points correspond to subjects drawn from the training partition. Moreover, if the subject is a true positive case of "Good QOL", then her score is represented as a point in the upper level (tr/TP), otherwise it appears as a point in the lower level (tr/TN). In the lower half of the same subplot, points are organised in the same way, except for the fact that they correspond to subjects drawn from the test partition (thus the prefix "te"). The matching scores depicted in the "Poor QOL" subplot are organised similarly.

Finally, it should be noted that this graph indicates poor model performance when there is a large overlap between the interval $[a, b]$ enclosing the matching scores from the true negative cases (te/TN), and the interval $[c, d]$ enclosing the matching scores from the true positive cases (te/TP). More precisely, the larger the value of $|[\,a, b\,] \cap [\,c, d\,]|$, the lower the model's performance for a given label j, where:

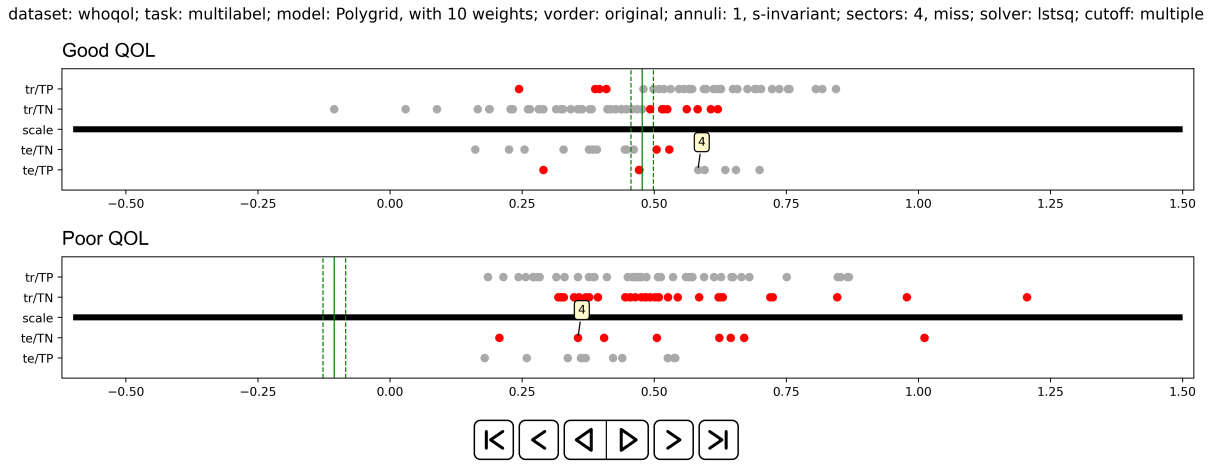
$$a, b := (min, max)\{\, \langle\, S_i, W_j \,\rangle : Y_{ij} = 0,\ i \in \text{test set}\,\},$$
$$c, d := (min, max)\{\, \langle\, S_i, W_j \,\rangle : Y_{ij} = 1,\ i \in \text{test set}\,\}.$$

The ideal condition is described by $[\,a, b\,] \cap [\,c, d\,] = \varnothing$, which implies the existence of a scalar $\nu_j \in [\,a, b\,] \cup [\,c, d\,]$ that separates the scores from positive and negative cases in the test set. The poor result in Figure 19 reflects the performance of the minimal Polygrid model, with $n_a = 1$. However, adding an extra annulus to this model substantially improves its performance[7], as shown in the second frame of that figure. However, as the subsequent frames illustrate, the performance does not respond monotonically to increments in n_{as}.

[6] Strictly speaking, we should have plotted the points corresponding to $\langle\, S_i, W_j \,\rangle \|W_j\|$, which gives W_j the role of unit of measure in this abstract ruler. However, it is convenient to strip away differences in size among the row vectors of W. For that purpose, we rescale each ruler with $\alpha_j = 1/\|W_j\|$, making $\alpha_j \langle\, S_i, W_j \,\rangle \|W_j\| = \langle\, S_i, W_j \,\rangle$. For an interactive visualisation of these ideas on the whoqol dataset with $d = 3$, please consult this Geogebra 3D applet (here).

[7] The accuracy of the model with a single annulus ($n_a=1$) is 0.5, and it increases to 0.9 after an extra annulus is added. For $n_a=3$, the accuracy increases to 0.95, and no misclassifications occur with $n_a=4$. These results refer to a random (but replicable) partitioning of the dataset.

Figure 19 – A scales diagram with matching scores from the whoqol dataset



Source: The author.

Legend: This diagram organises matching scores (points), in four levels: label, partition, ground case, and predicted case. **(1)** The first level corresponds to labels. The upper subplot corresponds to the "Good QOL" label, and the lower one to the "Poor QOL" label. The thick line that runs through each subplot depicts a compact set within the subspace spanned by the vector W_j. The graduation in the abscissa identifies the specific set being shown. **(2)** The second level encodes the allocation of subjects to the training or to the test partition. The scores from subjects in the training partition are plotted above the thick line, and the ones from subjects in the test partition appears below it. **(3)** The third level encodes the ground case: scores plotted in the "*/TN" rows correspond to true negative cases, and the scores that appear in the "*/TP" rows are true positive cases. **(4)** The fourth level encodes the predicted case. The solid green line represents a threshold value (see Algorithm 5). Ideally, true negative cases should appear to the left of this line, and true positive cases should appear to its right. Cases that fail to meet this condition are painted in red. Finally, the green dashed lines surrounding the threshold depict the granularity parameter described in Algorithm 5.

Notes: Click on the play button to start an animation showing scale diagrams for increasing number of annuli ($n_a = 1 \ldots 8$) based on the model previously described in Figure 13.

4.5 The interpretability of the Polygrid model

The Polygrid model generates highly faithful explanations. In principle, a person can replicate any prediction using only its respective explanation diagram, stripped of its output tags but with areas and weights annotated, printed on a sheet of paper. Intuitively, this seems to be a sufficient reason to state that the Polygrid model is highly transparent. However, the status of the Polygrid model with respect to interpretability is less clear. On what grounds can we claim that the Polygrid model is interpretable when the very notion of interpretability is still being debated? If interpretability is framed as a property, is it a matter of kind or degree? What are the classes of objects (e.g., models, model instances, predictions, explanations) to which interpretability can be ascribed and what is entailed by this? Is the interpretability of a model dependent on the properties of its inputs (e.g., data cases, hyperparameters) or somehow related with the properties of its outputs?

Answers to these questions must take into account two ideas. First, some consensus has emerged from the debate in machine learning about the facets of and how to assess interpretability, based on the analysis of the so-called "inherently interpretable models", such as linear regression and decision trees (Lipton, 2018; Rudin, 2019). Second, the quality of an explanation is a subjective notion in the sense that an explanation about an event E may be well received by person A and not so well by person B (Miller, 2019). These ideas are useful to show how distinct aspects of interpretability can be coordinated to justify the claim that Polygrid is an interpretable model. But before we develop this argument, we need to make our usage of some key terms explicit, in order to circumvent the tensions that emerge when using terms that have been inconsistently defined in the literature.

4.5.1 Preliminaries

Let's start by clarifying the distinction between the terms model and model instance. We stand by their definitions in Section 3.2.1: a model is an abstract structure that spans the hypothesis space $\mathcal{H}$, and a model instance is a hypothesis $h \in \mathcal{H}$ after its parameters have been adjusted to the training data. However, in the interest of precision, these definitions must be extended to include details that will allow us to speak of "model size".

Accordingly, we define a model as a quadruple (C, A, H, O) with a coordinator C, an architecture A, hyperparameters H, and an optimisation procedure O. The coordinator organises the operation of the model (e.g., it runs the sequence of steps needed to instantiate the model and fit it to the data, or to produce a prediction for an unseen input). The architecture A is an abstract forward computational graph, as illustrated in Figures 20a and 20c. A forward computational graph is a directed acyclic graph that represents a set of computations applied to the input data to produce an output. In such a graph, an edge represents a parameter (aka weight) and a node represents an intermediate variable and possibly any computations required to update that variable (Fang *et al.*, 2024). The coordinator instantiates the architecture A according to the dimensionality of the training data and any relevant hyperparameters in H (e.g., the number of hidden layers, the number of neurons per layer in a multilayer perceptron), and then tunes the weights of the instantiated graph h by calling the optimisation procedure O. The latter maximises the objective function $q(h, D)$ seen in Section 3.2.1, which depends on the input data and the weights of the instantiated graph h. Two instantiated graphs corresponding to the architectures in Figures 20a and 20c are illustrated in Figures 20b and 20d, respectively.

To compute a prediction for an unseen input X_i, the coordinator submits the input to the instantiated graph h, collects the results $h(X_i)$ and returns it to the caller[8]. To produce an explanation, the coordinator transforms the instantiated graph h, with the

[8] This assumes that the presentation of recommendations and explanations to the user is organised by services provided by the platform on which the recommendation model runs.

values of the nodes updated during the prediction stage, into some predefined format and returns it to the caller. This conceptual setup allows us to define fidelity as a relation between a model and the explanations it generates, independent of human judgement: an explanation is faithful if the forward computational graph of its corresponding prediction can be derived from it[9]. In addition, we define that a model is transparent to the extent it generates faithful explanations. In this framing, the Polygrid model is highly transparent.

4.5.2 Interpretability as a multidimensional property

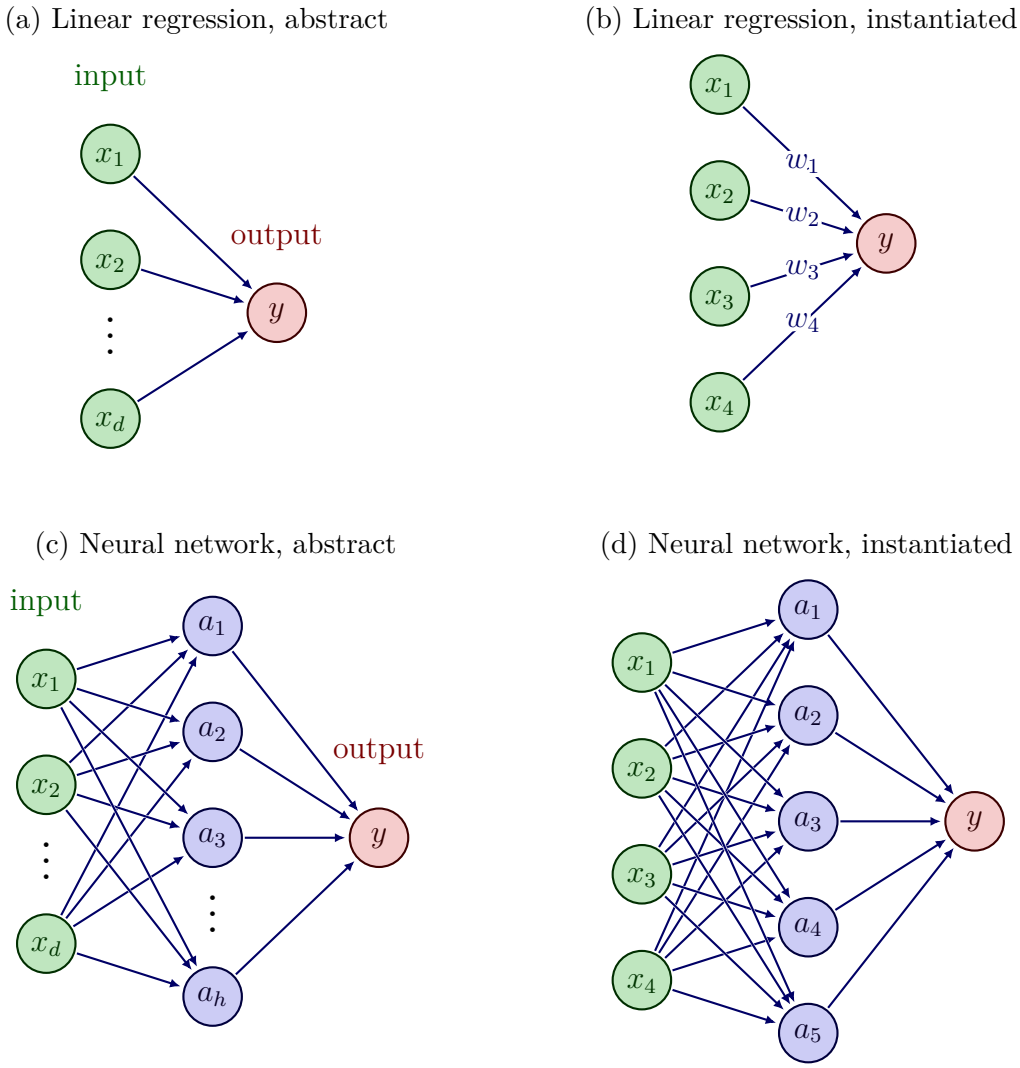
In the discussion that follows, we adopt an operational definition of interpretability described by Doshi-Velez and Kim (2017): interpretability is the measure of success participants achieve when asked to perform forward simulation tasks in an experimental context: "humans are presented with an explanation and an input, and must correctly simulate the model's output." This definition and its underlying methodology imply a graded notion of interpretability. In addition, note that it is consistent with methodological practice in visualisation research: the effectiveness of a visualisation in a decision-making task is measured by the success that participants achieve in an experimental setting (see Section 3.3). This is a fortunate coincidence because, in our case, the forward simulation tasks needed to assess model interpretability can be identified with the decision-making tasks needed to assess the effectiveness of the Polygrid explanation diagram.

It is reasonable to expect that any account of interpretability should agree with the idea that higher dimensional data are less conducive to interpretability than lower dimensional data. This expectation is reasonable because "interpretation" demands human cognition, which is a limited resource, and higher dimensional data ultimately implies higher cognitive demands[10]. Another characteristic of the data that is relevant to interpretability is meaningfulness. This is to say that being able to ascribe some meaning to each feature of the data is conducive to interpretability. Let's perform a thought experiment to bring this point home: imagine that a layperson is participating in a within-subjects study with two experimental conditions. In both, the participant is asked to perform five forward simulation tasks for a linear regression model. In the first condition, the subject is presented with data from the UCI Real State Valuation dataset, and is asked to predict the price per unit of area of a house. This dataset has six features, and all of them are meaningful to the general public (e.g. location, house age, distance to the nearest metro station).

[9] Note that, even in this user-independent framing, the degree of transparency can be modulated by factors that are unrelated to the model. For example, in our case, in which the explanations can be expressed as a spreadsheet, transparency can be modulated by the number of decimal places with which the values corresponding to weights and areas are stored in the spreadsheet.

[10] Limited in the sense that any task demands some degree of cognitive effort to be completed. For example, assume that a subject participating in a user study involving decision-making tasks knows how to solve them optimally. She may decide to match the cognitive demand, to pursue a less costly strategy (at the risk of lower performance), or even abandon the task.

Figure 20 – Examples of abstract and instantiated forward computational graphs



Source: The author.

Legend: A forward computational graph is directed acyclic graph that formalises a computation. (a) an abstract computational graph (aka an architecture) of a linear regression model. (b) the previous graph instantiated to fit a dataset with $d = 4$ features. Each edge stores a weight w_k and connects an input node x_k to the output node y. The output node y encodes the forward computation $y = \sum_k x_k w_k$, with $k \in 1 \dots d$. (c) an abstract computational graph of a multilayer perceptron with a single hidden layer. (d) the previous graph instantiated to fit a dataset with $d = 4$ features, and a hidden layer with $h = 5$ neurons. The output node y encodes the forward computation $y = \sum_j a_j w_j^{(2)}$, $j \in 1 \dots h$, with $a_j = \sigma(\sum_k x_k w_{jk}^{(1)})$, $k \in 1 \dots d$, and $\sigma(\cdot)$ being some sigmoid function. As mentioned earlier, a model is defined by a quadruple (C, A, H, O). In the training stage, the coordinator C instantiates the architecture A to fit the dimensionality d of the training data, and also any relevant hyperparameters H, such as the number of neurons in the hidden layer in Figure (d). The coordinator then applies the optimisation procedure O to tune the weights of the model. Once the training stage is completed, the coordinator produces a prediction by submitting unseen data X_i to the instantiated graph, which produces a forward computation y_i. The coordinator then collects the result y_i and includes in the output of the model, which may contain other information.

The explanation is provided as a computational graph printed on a sheet of paper, with variable names and weights updated, and formulas for each node described in the legend. This enables the participant to use her understanding of the problem domain in numerous creative ways, and some of these ways are useful in reviewing her answers before submission. In the second condition, the subject is presented with the same data, but the names of the features and any contextual elements that could hint at the domain problem are replaced with unrecognisable symbols (e.g., the house age is renamed as "0D280", and the price per unit area as "Y"). The participant cannot review her answers in light of her knowledge of the domain and is more prone to make mistakes because of a larger cognitive demand to process unfamiliar symbols. In principle, this blinding of the meaningful elements of the problem could lead to a lower average rate of success compared to the first condition of the study, which implies lower interpretability in the adopted methodological framework.

With respect to how a model contributes to interpretability, there are (at least) two relevant characteristics which are traceable to the model's architecture: scalability and meaning preservation. Consider the first experimental condition of our thought experiment. The linear regression model illustrated in Figure 20a takes up $d + 1$ weights, and the meaning of each weight is derived from the meaning of its corresponding feature (i.e., the rate of change of the target variable per unit change in the feature), except for the intercept. Now suppose that we replaced the linear regression model with the multilayer perceptron shown in Figure 20c with $h = 5$ neurons in the hidden layer. This instance would take up $5d + 11$ weights, and the architecture of the model makes it difficult to ascribe meaning to the learned weights because the user cannot ascribe meaning to the hidden nodes. This setting implies that the participant is subject to a higher cognitive demand because the instantiated graph is much larger, so there are more opportunities for mistakes and more incentive to defer the task, which could lead to a lower average rate of success in our hypothetical user study. Even if the linear model were replaced with a multilayer perceptron with $h = 2$ neurons in the hidden layer, which would result in a smaller difference in the size of the instances ($2d + 5$ weights), the participant's ability to use domain knowledge is reduced because the hidden nodes of the computational graph do not correspond to any concepts in the problem domain. In a sense, we are advancing an operational notion of meaning preservation as the ability of the study participant to ascribe meaning to different elements of the forward computational graph.

According to this analysis, it seems to be more fruitful to conceive of interpretability as a multidimensional property that emerges from the interaction among multiple factors[11].

[11] Here, the term "interaction" is being used in the sense it is employed in discussions about perspectivism. As an example from Giere (2006) goes, the colour of a sweater is not an intrinsic property of that object. Many factors determine our perception of colour: characteristics of the sources of light and their placement in the environment, characteristics of the fabric, as well as conditions of the viewer that may affect her perception (e.g., colour blindness).

However, it must be noted that high levels of interpretability can only be achieved in settings where the model is transparent and preserves the meaning of the input data.

4.5.3 A defence of Polygrid's interpretability

Assuming the reader agrees with our tactical choice for an operational definition of interpretability and with our analysis about its contributing factors, we now proceed to justify the claim that Polygrid is an interpretable model on our conceptual framing.

Let's start with the data. As seen in Section 2.4, psychometric data are typically low dimensional and meaningful to our target user: the attending care professional was trained to use the instrument for CGA. Since both of these characteristics are conducive to interpretability in our framing, it remains to be shown that Polygrid preserves the meaning of the data. We argue that it does on the basis of three premises about how Polygrid transforms the input data into predictions, and how the data are depicted in the explanation diagram. The argument goes like this:

(P1) The assessment polygon is a visual analogue of the measurand.
(P2) The measurand is meaningful to the user.

(C1) The assessment polygon is meaningful to the user.

(P3) The transformation of the assessment scores into the assessment polygon preserves the meaning of the assessment scores.
(P4) The assessment scores are meaningful to the user.

(C1) The assessment polygon is meaningful to the user.

(P5) The transformation of the assessment polygon into the matching polygon preserves the meanings of the assessment polygon.
(C1) The assessment polygon is meaningful to the user.

(C2) The matching polygon is meaningful to the user.

The premises P2 and P4 are true because the user has been trained to use the instrument. We argue that P1 is true due to the relationship between the measurand and the area of the assessment polygon. Recall from Section 2.3 that, for the instruments designed in the factor-analytic tradition, the standing of an individual on the latent variable is often estimated as a weighted sum between the domain scores and the instrument's loadings. For instruments with a scoring procedure that produces a sum-score using Equation 2.6, it can be shown that the area of the assessment polygon induces an ordering of subjects on the measurand that is equal to the ordering induced by the sum-scores under certain assumptions. This result warrants identifying the meaning of the measurand with the meaning of area of the assessment polygon, for the purpose of ordering people on the

latent variable. This relationship is fully detailed in Appendix A, but we present a cut version in the following. Under the assumptions:

(A1) $\mathring{x}_{ik} = \lambda_k \eta_i + \epsilon_{ik}$ (a structural equation from a congeneric factor model, Eq. 2.1);

(A2) $\eta_i > 0$, for all $i = 0 \ldots m-1$ (the latent variable tracks a capacity of individuals);

(A3) $\lambda_k > 0$, for all $k = 0 \ldots d-1$ (items correlate positively with the latent variable);

(A4) $\epsilon_{ik} = 0$ (the measurement error is negligible);

(A5) $\max\{\mathring{X}_{:k}\} = C$ for all $k \in 0 \ldots d-1, C > 0$ (all subscales have the same range);

whenever $\eta_a > \eta_b$, then it must be the case that:

$$\eta_a > \eta_b \overset{A3}{\implies} \implies \lambda_k \eta_a > \lambda_k \eta_b \overset{A1,A4}{\implies} \mathring{x}_{ak} > \mathring{x}_{bk} \overset{A5,A2}{\implies} x_{ak} > x_{bk} > 0 \ (\forall k \in 0 \ldots d-1)$$

$$(x_{ak} > x_{bk}) \wedge (x_{a,k+1} > x_{b,k+1}) \implies x_{ak} x_{a,k+1} > x_{bk} x_{b,k+1} \ (\forall k \in 0 \ldots d-1)$$

$$\text{take } 2\nu = sin\Big(\frac{2\pi}{d}\Big) > 0,\ \nu x_{ak} x_{a,k+1} > \nu x_{bk} x_{b,k+1} \implies \sigma_{ak} > \sigma_{bk} \ (\forall k \in 0 \ldots d-1)$$

$$\sigma_{ak} > \sigma_{bk} \implies \sum_k \sigma_{ak} > \sum_k \sigma_{bk} \implies \sigma_a > \sigma_b.$$

The statement $\eta_a > \eta_b$ speaks about two individuals that have distinct levels of the measurand η (e.g., quality of life), with person A having a higher position on the latent variable than person B. Accordingly, the implied statement $\sigma_a > \sigma_b$ indicates that the area of the assessment polygon for person A is greater than that of person B. In these equations, σ_{ak} corresponds to the area of the polygonal shape formed by two consecutive domain scores, namely $\mu(\triangle(0,\, x_{ak}\,\zeta_k,\, x_{a,k+1}\,\zeta_{k+1}))$, indices modulo d as usual[12].

Based on this result, we argue that premise P3 is true because the user can ascribe meaning to the elements of the computational graph that correspond to mapping assessment scores to the assessment polygon. For example, the forward computational graph shown in Figure 26b was extracted from the explanation diagram in Figure 15a. In this computational graph, the elements $T_0, \ldots, T_3$ correspond each to a triangle formed by consecutive vertices of the assessment polygon (and the origin). Thus, since the area of the assessment polygon corresponds to the measurand, the meaning of a node T_q is an additive component of the measurand that is determined by two domain scores.

Similarly, we argue that premise P5 is also true because the user can ascribe meaning to the remainder of the computational graph. In fact, all nodes represented in blue perform some function related to feature extraction, and this is precisely because each node represents some additive component of the area of the assessment polygon. The

[12] The expression $\triangle(z_0, \ldots, z_{d-1})$ indicates that the tuple $(z_0, \ldots, z_{d-1})$, taken as the closed polygonal chain $(z_0, \ldots, z_{d-1}, z_0)$, specifies a solid polygonal shape.

edges connecting the two outer layers derive their meaning from the nodes they connect: they represent the rate of change of the target variable per unit change in the feature, as in linear regression. Finally, the remainder of the edges do not represent learned weights: they are constants. As shown, the Polygrid model preserves the meaning of the input data and, as seen earlier, it is transparent, and both properties are conducive to interpretability.

4.6 Summary and closing remarks

This chapter introduced the Polygrid model and detailed how it performs multilabel classification and label ranking tasks. The two seminal ideas in the development of this model can be traced to the literature on gerontology: the use of psychometric instruments to standardise the comprehensive assessment of patients (CGA), and the recurrent and independent calls from authors to represent the assessment results as a radar chart. Then, two key insights have driven the development of the model. The first connects psychometric instruments with the radar chart showing a patient's assessment: the congeneric model, the workhorse of psychometric instruments (in the factor-analytic tradition), constrains some geometric features of the polygon in that radar chart. This insight allowed us to ascribe meaning to the area of the polygon that depicts the results of the patient's assessment, which is a core element of our argument for the interpretability of the model. The second insight is a change in perspective, shifting from "the radar chart as a visualisation" to "the radar chart as an image". In a sense, the partitioning of the unit disc and the decomposition of the assessment polygon, two important steps in the Polygrid's learning pipeline, functionally resemble the initial steps of a (primitive) computer vision pipeline: the image segmentation and feature extraction steps that convert an image into a feature vector. This insight gave us the ability to project the assessment data into higher dimensional spaces and explore techniques in collaborative filtering and multilabel ranking to devise the learning pipelines.

Parallel to this, the requirement for interpretability imposes the constraints needed to align the model's architecture and the design of the explanation diagram. To preserve meaning, the architecture has a single layer with learned weights, as in linear regression. All remaining layers perform feature extraction, and the extracted features represent additive components of the measurand (i.e., the latent variable that the instrument measures). Regarding the explanation diagram, the idea of representing interventions in the same space we represent assessments (the assignment chart) comes from collaborative filtering. The missing element, the matching chart, resulted from the idea of depicting the inner product $\langle\, s, W_j^{\mathsf{T}} \,\rangle$ as the weighted area of the assessment polygon. Finally, to ensure transparency, the forward computational graph is visually encoded in the diagram's design.

The resulting model is tested in the next chapters. Evidence in support of Polygrid's learnability is presented in Chapter 5, which reports the results of an offline performance evaluation, and Chapter 6 presents evidence of its interpretability from a user study.

5 AN OFFLINE EVALUATION OF THE POLYGRID MODEL

In the tradition of recommender systems research (and machine learning in general), the first test a new model must pass on its way to field deployment is an offline evaluation. The idea is to simulate a real-world setting and assess the degree of success the model achieves in the tasks it will perform once deployed. In this chapter, we report on the results of an offline evaluation of the Polygrid model. We describe its performance against a number of datasets from the healthcare domain. These datasets contain health-related assessments using standardised instruments (i.e., psychometric data) and assignments that can be cast either as multilabel or label ranking assignments. Once Polygrid has been fit to the data, several metrics are calculated and compared with a set of alternative models, whose performance were subjected to unusually strict conditions. They are unusual because they include controlling for model size[1]. But why should we control for model size? The design of the evaluation was crafted with three goals in mind:

1. Considering multilabel classification and label ranking as relevant tasks, we want to show how competitive the Polygrid model is compared to top-performing models.

2. If an alternative model achieves a degree of performance that is markedly superior to that of Polygrid, we want to answer the question: **What did it do differently?** In other words, we want to be able to ascribe an observed difference in performance to some observable difference in the constitutive characteristics of the models (e.g., differences in their forward computational graphs). The aim is to identify ideas to improve the performance of the Polygrid model to be explored in future work.

3. The Polygrid model has several hyperparameters. Some determine its size (e.g., the number of annuli and the number of sectors per domain), and others determine how the feature space is partitioned (e.g., how sectors and annuli cover the unit disc). The results should provide us with some hints about how to choose the hyperparameters of the model considering the characteristics of the target dataset.

As the reader may have noticed, the first two goals are in tension. The first goal aims to rank the new model among top performers based on its performance on a collection of datasets. This corresponds to the standard practice of evaluating a new model in a mature domain (Demšar, 2006) and, in such a setting, the size of a model is usually not taken into account directly. In contrast, the second goal is to identify potential improvements for the

[1] The term "model size" refers to the average number of weights taken up by instances of a model when trained on some dataset. Thus, it is a context-dependent notion. The concept of weight was presented in Section 4.5.1 as a property of an edge of a computational graph.

Polygrid model. In this context, an improvement is an answer to the posed question ("What did it do differently?"). However, we are not interested in answers that ascribe an observed difference in performance to a difference in size of the models being compared. This is because interpretability deteriorates as the model size increases and, since our application domain values higher degrees of interpretability, it seems reasonable to constrain the size of the models during evaluation. We assume that the Polygrid model is interpretable up to some size and constrain the evaluation to this limit. This assumption will be tested in Chapter 6, at least in part. As a consequence, our evaluation trades off the two goals at a cost: the rankings reported in this chapter cannot be interpreted as unconditional statements, such as "the performance of the model M_1 dominated the performance of M_2." Instead, a statement like $M_1 \succ M_2$ must be read as "for models of a certain size, the performance of M_1 dominated that of M_2."

Another reason why this evaluation departs from the standard design is that the application of recommender systems to healthcare is far from being a mature domain, as discussed in Section 3.1. In fact, the application of recommender systems in gerontological care is in its infancy: there are no standard tasks or metrics, let alone public benchmark datasets. This means that pursuing the first goal to the detriment of the others would lead us to less useful results. For example, knowing that a multilayer perceptron with a single hidden layer and about 700 weights dominates the performance of Polygrid in a certain healthcare dataset. Arguably, this result does not provide useful guidance to model selection, given that it would deliver poorly on interpretability because of its size.

The remainder of this chapter is organised as follows. The design of this evaluation is detailed in Section 5.1. It includes an explanation of how the three evaluation goals are met. The results of the evaluation are presented in Section 5.2 and discussed in Section 5.3. Since controlling for model size in comparative evaluations is unusual, concerns about its fairness are also discussed. Section 5.3.5 compiles some lessons about the selection of hyperparameters for the Polygrid model. Finally, Section 5.4 concludes.

5.1 The design of the offline performance evaluation

In a nutshell, this is an offline evaluation in the tradition of recommender systems' research. It assesses nine models across 15 datasets using nine distinct metrics. We follow a standard protocol to rank the models according to their performance (Demšar, 2006), but the protocol is adapted to tackle limitations of the statistical test that arise because the number of datasets used as a benchmark is too small.

In the following, we detail how the datasets were collected and curated, how the models and relevant metrics were selected, and how the evaluation process explores the Polygrid's configuration space, generates the performance data, and manages the size of the models. Finally, the adaptation of the statistical analysis is justified and explained.

5.1.1 Datasets

Ideally, all datasets included in this evaluation should be similar to those collected by research initiatives implementing the WHO/ICOPE approach to reorient primary care services for older people (Tavassoli *et al.*, 2022; Ferriolli *et al.*, 2024). These initiatives stand out for their emphasis on the quality of the collected data: a large body of health professionals is trained to assess members of a well-defined target population using a standardised instrument, a large number of members of the target population is engaged in the study, and all recommendations made for each participant by their attending primary care professional are recorded. Unfortunately, we were unable to access these datasets, either due to data protection barriers or because the study is still ongoing. To overcome this limitation, we secured access to three datasets containing health-related assessments:

- WHOQOL: this dataset was collected by researchers affiliated with the Department of Gerontology at the Federal University of São Carlos (Brazil) for a study reported in Lorenzi *et al.* (2022). It contains quality of life assessments of 100 older individuals (50 years or older, average 67 years, 86% female) who participated in U3A activities promoted by the FESC institution in October 2019. Assessments were conducted using the WHOQOL-BREF instrument (Orley *et al.*, 1996), but no recommendations or referrals were systematically recorded. In other words, using the nomenclature defined in Section 4.1, this dataset contains only the description matrix, $\mathring{D} = (\mathring{X}, -)$.

- AMPIAB: this dataset was collected by researchers affiliated with the University of São Paulo (Brazil) for a study reported in Andrade *et al.* (2020). It contains assessments of the health deficits of 510 people (60 years or older, average 76 years, 78% women) who were home-assisted by the Elderly Caregiver Programme promoted by the Health Department of the city of São Paulo. Assessments were conducted in November 2018 using the AMPI/AB instrument (Andrade, 2019), but referrals were recorded for 128 patients only.

- ELSIO1: this dataset was derived from a publicly available dataset collected by researchers that contributed to the Brazilian Longitudinal Study of Aging (ELSI-Brazil), led by Fiocruz (Lima-Costa *et al.*, 2018). The researchers conducted interviews and collected physical measurements of 9,412 individuals in a nationally representative sample of community-dwelling adults aged 50 years or more over the years 2015-16. The original data were transformed into assessments of intrinsic capacity (WHOIC instrument) following a method employed in Aliberti *et al.* (2022), resulting in 7,175 derived assessments. For reasons related to the computational cost of the Polygrid model, we decided to use a small sample of the data (718 out of 7,175 derived assessments). Similar to the WHOQOL dataset, this contains only the feature data.

In summary, these datasets contain health assessment data of distinct Brazilian populations (people participating in U3A activities in São Carlos, people receiving health care at home in São Paulo, and community dwellers nationwide). Data were collected with different instruments (WHOQOL and WHOIC capture health capacity, and AMPIA/AB captures health deficits). These instruments were developed in the factor-analytic tradition. Referrals are mostly absent and, to address this issue, we augmented the datasets with synthetic assignments. The latter were generated based on a core principle in healthcare that asserts that individuals with similar needs should receive similar care (Varkey, 2020). Furthermore, these assignments were constrained so that they reproduce key distributional properties that were observed in two real-world datasets, as explained next.

Based on these ideas, we created five versions of each dataset listed above, as summarised in Table 12. Each dataset has a single multiclass version. The assignments in the "whoqol" dataset stratify the sample into individuals with good and poor quality of life, according to a criterion proposed by Silva *et al.* (2014) — individuals with a score equal to or greater than 60 are assigned to the "Good QoL" class. The "ampiab" dataset has assignments based on the criterion described in Andrade *et al.* (2020), which classifies individuals as healthy, pre-frail, and frail. And the assignments in the "elsio1" dataset were created by computing the Katz index[2] for each individual, as in Aliberti *et al.* (2022).

In addition, each dataset has two multilabel variants, whose identifiers end with "-ml-11" or "-ml-22". With the exception of the "ampiab-ml-11" dataset, their assignments were created by applying fuzzy clustering to the assessment data, subject to a constraint on the λ-cut to reproduce distributional properties seen in two studies mentioned before:

- A subset of the AMPIAB dataset in which instances have informed referrals (Andrade *et al.*, 2020). This sample is taken as the "ampiab-ml-11" dataset, which has 128 instances and 11 labels (referrals). The cardinality observed in this sample is 1.08 (i.e., each individual is assigned to 1.08 labels on average). This statistic was used to calibrate the λ-cut when creating the assignments for the other "*-ml-11" datasets.
- Although we did not have access to the dataset collected for the study reported in Tavassoli *et al.* (2022), we used the summary data in Table 3 of that work to infer that the dataset has 22 labels and a cardinality around 4.54. This statistic was used to calibrate the constraint when creating assignments for the "*-ml-22" datasets.

In other words, a raw dataset $\mathring{D}$ was transformed into a multilabel dataset D by:

$$\underset{\substack{\downarrow\\ \textit{original}\\ \textit{dataset}}}{\mathring{D}} = (\mathring{X}, -) \mapsto (\underset{\substack{\downarrow\\ \textit{scaled to}\\ \textit{the unit}\\ \textit{hypercube}}}{X}, -) \mapsto (X, \underset{\substack{\downarrow\\ \textit{constrained}\\ \textit{fuzzy}\\ \textit{clustering}}}{u(X)}) = (X, Y) = \underset{\substack{\downarrow\\ \textit{multilabel}\\ \textit{dataset}}}{D}. \tag{5.1}$$

[2] The Katz index summarises the ability of a person to perform activities of daily living, such as bathing, dressing, going to toilet, transferring, continence, and feeding (Katz *et al.*, 1963).

Finally, each dataset also has two label-ranking variants, which correspond to datasets ending with "-lr-11" or "-lr-22". The assignments were created using the method described for the multilabel datasets, with the additional step of using fuzzy membership scores to induce the ordering of the labels in each individual assignment (Ross, 2010).

Table 12 – Characteristics of the datasets used in the offline evaluation

dataset	assignment	#inst	#ft	#lb	card.	imb.	#ls	#sls	max l/i
whoqol	multiclass	100	4	2	1.00	0.31	2	0	1
whoqol-ml-11	multilabel	100	4	11	1.09	0.87	16	4	2
whoqol-ml-22	multilabel	100	4	22	4.54	0.72	77	63	8
whoqol-lr-11	label ranking	100	4	11	1.09	0.87	18	7	2
whoqol-lr-22	label ranking	100	4	22	4.54	0.72	92	86	8
ampiab	multiclass	510	5	3	1.00	0.64	3	0	1
ampiab-ml-11	multilabel	128	5	11	1.08	0.99	15	8	2
ampiab-ml-22	multilabel	510	5	22	4.54	0.61	215	113	13
ampiab-lr-11	label ranking	510	5	11	1.08	0.56	25	6	2
ampiab-lr-22	label ranking	510	5	22	4.54	0.61	259	168	13
elsio1	multiclass	718	5	6	1.00	1.00	6	0	1
elsio1-ml-11	multilabel	718	5	11	1.08	0.49	20	2	2
elsio1-ml-22	multilabel	718	5	22	4.54	0.58	234	86	10
elsio1-lr-11	label ranking	718	5	11	1.08	0.49	25	3	2
elsio1-lr-22	label ranking	718	5	22	4.54	0.58	431	299	10

Source: The author.

Note: Column "#inst" shows the number of instances, "#ft" of features, and "#lb" of labels. "card." stands for cardinality, "imb." for imbalance, "#ls" is the number of labelsets and "#sls" the number of single labelsets. The maximum number of labels per individual in a dataset is shown in "max l/i". The imbalance score is computed with Equation 3.2.

5.1.2 Alternative models

The following models were chosen because they appear at the top of the rankings published in the reference studies that were described in Section 3.2:

- Top-performing models for multilabel classification: in the rankings published in Bogatinovski *et al.* (2022), the Binary Relevance with Random Forests of Decision Trees (BRRF) is the best performing model across multiple metrics among models that follow a problem transformation approach, while the Random Forest of Decision Trees (RF) is among the best models in the ensemble of adapted algorithms approach[3].

[3] In Bogatinovski *et al.* (2022), the Binary Relevance with Random Forests of Decision Trees is referred to as RFDTBR, and the Random Forest of Predicting Clustering Trees is referred to as RFPCT, but we refer to them as BRRF and RF, respectively. The latter is a Breiman's Random Forest, and since a Predictive Clustering Tree is induced with the ID3 algorithm, we assume it is similar to a CART tree. This hypothesis is supported by the results in Table 13.

- Top-performing models for label ranking: according to a ranking published in the supplementary material of Fotakis, Kalavasis and Psaroudaki (2022), the Random Forest (RF) model is competitive with the best performing models when evaluated against the semi-synthetic benchmark datasets, besides being relatively simple.

To assess the degree to which the implementations we adopted are equivalent to the ones used in the studies above[4], their evaluation was replicated for some benchmark datasets and the results compared[5]. The Foodtruck and Water datasets from the Cometa and KDIS repositories were selected as baselines for the multilabel classification task, and the Vowel dataset from the Paderborn repository was picked as a baseline for the label ranking task. These datasets were chosen because their number of attributes and classes best matched our healthcare datasets (that is, having about 11 labels and low-dimensional feature data). The results are shown in Table 13, and indicate that the performance of both models was more closely replicated for the Water dataset than for the Foodtruck dataset, and also indicate a large difference with results reported for the Vowel dataset.

Table 13 – Results from replicating the evaluation of top-performing models for multilabel classification and label ranking tasks

				BRRF				RF			
Dataset	**#ft**	**#lb**	**metric**	**base**	**lb**	**ub**	**outc**	**base**	**lb**	**ub**	**outc**
Foodtruck	11	12	accuracy	0.123	0.105	0.167	success	0.110	0.135	0.185	↑
Foodtruck	11	12	hamming	0.156	0.179	0.206	↑	0.155	0.174	0.189	↑
Foodtruck	11	12	f1.micro	0.587	0.455	0.495	↓	0.600	0.480	0.513	↓
Foodtruck	11	12	f1.macro	0.253	0.203	0.260	success	0.213	0.181	0.248	success
Water	16	14	accuracy	0.026	0.014	0.028	success	0.011	0.012	0.025	near
Water	16	14	hamming	0.281	0.267	0.281	success	0.280	0.267	0.283	success
Water	16	14	f1.micro	0.626	0.603	0.624	near	0.631	0.614	0.630	near
Water	16	14	f1.macro	0.573	0.551	0.581	success	0.568	0.561	0.583	success
Vowel	10	11	Kendall's τ	-	-	-	-	0.67	0.257	0.291	↓

Source: The author.

Note: The column "#ft" shows the number of features, "#lb" the number of labels, "base" shows the baseline from the reference study, "lb" and "ub" are the lower and upper bounds of the replicated measurement. In "outc" (outcome), "success" indicates a measurement compatible with the baseline, "near" indicates a near hit, and ↓↑ mark deviations from the baseline. Bootstrapped confidence intervals were computed on samples of size 10, at the 5% level adjusted with Bonferroni's method. RF has 100 DTs of unrestrained depth.

We impose the use of the CART Decision Tree (DT) as base learning model on BRRF and RF. Also, we include the DT model among the alternative models, as well as the Binary Relevance with Decision Trees (BRDT). The aim is to allows us to isolate the

[4] We use implementation from the scikit-learn and scikit-multilearn-ng packages.

[5] We adopt the definition of replicability being advanced by the ACM, see their website.

contribution of the binary relevance strategy from that of the bagging strategy of Random Forests[6] by comparing the performance of BRRF and RF with that of BRDT and DT.

The following models were also included: Least Squares Linear Regression (Linear), Ridge Regression (Ridge), and a Multilayer Perceptron with a single hidden layer and sigmoid activation (MLP). These models were included because of their structural similarity with the Polygrid model. For example, a Polygrid model with a single annulus and a single sector per domain, Polygrid$(1, 1)$, will take up as many parameters as a Linear or Ridge model. In this circumstance, differences in performance can be precisely determined. More specifically, the difference in performance between a Polygrid$(1, 1)$ model and a Linear or Ridge model can be solely ascribed to the partitioning of the feature space[7]. The MLP model was included because it generally projects the input data to a higher-dimensional space, similar to what Polygrid does. Accordingly, if an MLP instance and a Polygrid instance have the same size but there is a substantial difference in performance, then the difference can be ascribed to differences in how these models project the input data to the feature space and differences in the dimensionality of the target feature space. Finally, a Random model was included to provide us with intuitive lower bounds for sanity checking.

5.1.3 Relevant metrics

To evaluate models in multilabel classification tasks, we selected the following subset of the metrics reported in Bogatinovski *et al.* (2022): subset accuracy (which is referred to as accuracy for simplicity), micro and macro averaged F1 scores (f1.micro and f1.macro), label-weighted F1 score (f1.weigh), and the Hamming loss (hammingl). The accuracy metric indicates the share of predicted labelsets that exactly match the ground truth, f1.micro and hamming loss measure success at the level of the cells of the "prediction table", and the difference between f1.macro and f1.weigh indicates the impact of the imbalance of the dataset on model learning. In addition, we also implemented a set-based Jaccard similarity metric that is used to assess the degree to which the model can separate true positive cases from true negative cases, as illustrated in Figure 19.

Regarding model evaluation in label ranking tasks, we followed Fotakis, Kalavasis and Psaroudaki (2022) and selected the traditional Kendall's tau (ktau, Equation 3.5) metric. Moreover, we implemented a metric that is equivalent to subset accuracy, which we called lracc, that computes the fraction of predicted label rankings that exactly match the ground truth. We also implemented a Hamming loss equivalent metric, called lrloss, that computes the fraction of predicted label/rank that does not match the ground truth.

---

[6] The bagging strategy used in RF consists in (a) ignoring some less relevant features when looking for the best split, and (b) using a distinct random subset of the training data to fit each individual tree in the forest.

[7] For example, setting sector type to "miss" in Polygrid$(1, 1)$, as in Figure 13, replaces the attributes by their interactions in a linear regression: $(x_0, x_1, x_2, x_3) \to (x_0x_1, x_1x_2, x_2x_3, x_3x_0)$.

5.1.4 Evaluating the Polygrid model

The first stage of the evaluation consists of a search for the best Polygrid config for each dataset and metric. In this context, a config consists of a tuple indicating the number of sectors per domain ($1 \dots 3$), the number of annuli ($1 \dots 8$), the order of the vertices (3 options), the type of sector (2 options), the type of annuli (3 options), the type of solver (4 options), and the cutoff scheme (2 options), totalling 3,456 configs. Each config is evaluated $ss = 50$ times, producing samples of this size for every (dataset, config, metric)-triple. Therefore, we obtained 186,624 samples for multiclass and multilabel datasets, and 62,208 samples for label ranking datasets, resulting in 248,832 samples in total[8]. The sample size ss was empirically chosen so that the evaluation could be repeated on a different computing infrastructure with high consistency. The ranges for the number of sectors per domain and the number of annuli were chosen so that the workload demanded by the evaluation could be executed within a couple of weeks. Finally, this process is specified in Algorithm 10, and the steps of the algorithm are detailed in Table 14.

Algorithm 10: Perform an offline performance evaluation of the Polygrid model

Data: m-array *datasets*, n-array *models*, o-array *configs*, fn *metrics*, int ss
Result: (m, n, o)-array *results*, hash table *best_configs*

Function *eval-Polygrid*(*datasets*, *models*, *configs*, *metrics*, ss) **is**
  $j' \leftarrow index(\texttt{Polygrid}, models)$;
  for $i \in 0 \dots m-1$ **do**
    $(X, Y, U, task) \leftarrow load(datasets[i])$;
    $model\_type \leftarrow models[j']$;
    for $k \in 0 \dots o-1$ **do**
      $config \leftarrow configs[k]$;
      $L \leftarrow list()$;
      $sizes \leftarrow list()$;
      for $r \in 0 \dots ss-1$ **do**
        $(X_{tr}, Y_{tr}, U_{tr}, X_{te}, Y_{te}, U_{te}) \leftarrow split(X, Y, U, task)$;
        $model \leftarrow instantiate(model\_type, config, sizes)$;
        $model.fit(X_{tr}, Y_{tr}, U_{tr}, task)$;
        $\hat{Y}_{te} \leftarrow model.predict(X_{te})$;
        $L.append(evaluate(Y_{te}, \hat{Y}_{te}, metrics, task))$;
        $sizes.append(model.size)$;
      end
      $results[i, j', k] = summarise(L, sizes)$;
    end
  end
  $best\_configs = find\_best\_configs(results[:, j', :])$;
  return $(results, best\_configs)$;
end

[8] Details: 248,832 = 186,624 + 62,208, with 186,624 = 9 datasets × 6 metrics × 3,456 configs, and 62,208 = 6 datasets × 3 metrics × 3,456 configs.

Table 14 – A detailed description of the steps of Algorithm 10

Row	Description
2	Sets j' to the index of the Polygrid model in the *models* array.
4	Loads a dataset: X holds assessment data, Y holds assignments, U is a membership matrix, and *task* indicates either "multilabel classification" (MLC) or "label ranking" (LR). When *task* is MLC, then $U_{ij} \leftarrow Y_{ij} / \sum_k Y_{ik}$, meaning that all labels assigned to an assessment are equally relevant. In contrast, when *task* is LR, then $U \leftarrow logranks(Y)$, which applies the Equation 4.1 followed by row normalisation. This makes the labels at the head of a ranking more relevant than the ones at its tail.
5,7	Sets *model_type* to Polygrid, and selects the next $config$ to be assessed.
8,9	Sets L and *sizes* to empty lists.
11	Splits the dataset using a 80 : 20 proportion. The splitting algorithm computes the distribution of instances per label in the dataset, and uses this information to select the random partitioning of the dataset that most closely matches that distribution. The random partitioning is repeated a large number of times, for example, 10,000 times.
12	Instantiates a Polygrid model with the current $config$. The history of sizes of previously created instances, held in *sizes*, is ignored here, but this will play a key role in the evaluation of the alternative models.
13	Fits the Polygrid instance to the training data.
14	Submits the test data to the now trained Polygrid instance.
15	Evaluates the performance of the Polygrid instance across the metrics that are appropriate for the current *task*, and stores the results in the list L. After the loop started in row 10 completes, this list will contain *ss* items.
16	Determines the size of the current Polygrid instance and updates the history in *sizes*.
18	Uses the sample of performance data in L to compute a bootstrapped confidence interval for each metric that was assessed. The resulting statistics are stored as a dictionary in the *results* tensor, whose dimensions are indexed as (dataset, model, config). After the loop initiated in row 3 completes, *results* holds the evaluations for Polygrid across all datasets and configs.
21	Visits the lateral slice of the *results* tensor holding the results of the evaluation of Polygrid and determines the (ID of the) best configs for each dataset and metric.
22	Returns the *results* tensor (which only holds performance data for Polygrid) and the best Polygrid configs for each dataset and metric.

Source: The author.

5.1.5 Evaluating the alternative models

The second stage of the evaluation determines the performance of the alternative models in the best configs found in the previous stage. As mentioned earlier, a size constraint is placed on the instances of the alternative models. For example, the best Polygrid config found in the previous stage for the whoqol dataset ($d = 4$ features, $n = 2$ labels), based on the accuracy metric, was ($nspd = 1$, $na = 2$, vorder: averages, annulus: r-invariant, sector: cover, solver: ridge, cutoff: single). In this config, a Polygrid instance takes up 18 weights. This means that, when fitting the MLP model for this config, the number of neurons in the hidden layer is calculated so the number of weights taken up by the instance is close to 18. The architecture of the MLP model imposes $h = (target - n)/(d + n + 1) = (18 - 2)/(4 + 2 + 1) = 16/7 = 2.3$, with h being the number of neurons. Since h must be an integer, the strategy is to round off to the next integer ($h = 3$) and track deviations from the target number of weights so that the mean approaches the target value. If we set the number of repetitions to $ss = 10$, the strategy

will produce MLP instances with the following sizes: $[23, 9, 23, 16, 23, 9, 23, 16, 23, 9]$, which averages to 17.4. In this setting, an MLP instance with a single hidden neuron takes up 9 weights, with 2 neurons 16 weights, and with 3 neurons 23 weights. For DT and BRDT models, the average size of their instances is controlled by limiting the depth of their trees[9]. The same strategy applies to RF and BRRF instances, which have an additional limit on the number of trees per instance. For RF instances, the maximum number of trees is set to the number of existing labels. This allows for a structural comparison between RF and BRDT models, as both can grow the same number of trees. In the case of BRRF, each component RF is allowed to grow up to 10 trees. In contrast, the number of weights in instances of Linear and Ridge models is determined solely by the training data and, therefore, no strategy can be enforced to control their sizes. Finally, this evaluation stage is specified in Algorithm 11, and the steps of the algorithm are detailed in Table 15.

Algorithm 11: Performs a comparative evaluation of the alternative models

Data: m-array *datasets*, n-array *models*, o-array *configs*, hash table *best_configs*, (m, n, o)-array *results*, fn *metrics*, int *ss*
Result: (m, n, o)-array *results* (updated)

```
1  Function
     eval-Others(datasets, models, configs, best_configs, results, metrics, ss) is
2      j' ← index(Polygrid, models) ;
3      for i ∈ 0...m − 1 do
4          (X, Y, U, task) ← load(datasets[i]) ;
5          for metric ∈ metrics(task) do
6              k' ← best_configs[i][metric] ;
7              config ← configs[k'] ;
8              for j ∈ 0...n − 1, j ≠ j' do
9                  model_type ← models[j] ;
10                 L ← list() ;
11                 sizes ← list() ;
12                 for r ∈ 0...ss − 1 do
13                     (X_tr, Y_tr, U_tr, X_te, Y_te, U_te) ← split(X, Y, U, task) ;
14                     model ← instantiate(model_type, config, sizes) ;
15                     model.fit(X_tr, Y_tr, U_tr, task) ;
16                     Ŷ_te ← model.predict(X_te) ;
17                     L.append(evaluate(Y_te, Ŷ_te, metric, task)) ;
18                     sizes.append(model.size) ;
19                 end
20                 results[i, j, k'] = summarise(L, sizes) ;
21             end
22         end
23     end
24     return results ;
25 end
```

[9] A split node has two parameters (feature and threshold), whereas a leaf node has just one.

Table 15 – A detailed description of the steps of Algorithm 11

Row	Description
2	Sets j' to the index of the Polygrid model in *models* array.
4	Loads a dataset: X holds feature data, Y holds assignments, U is a membership matrix based on Y, and *task* indicates either "multilabel classification" (MLC) or "label ranking" (LR).
6,7*	Sets k' to the index of the best Polygrid config found in the previous stage for the i-th dataset and a particular *metric*. Also stores the tuple with the config specification in $config$.
9*	Selects the next *model_type* to be assessed.
10,11	Sets L and *sizes* to empty lists.
13	Splits the dataset using an $80:20$ proportion.
14*	Instantiates the alternative model being assessed using the current $config$ and *sizes*. The $config$ is used to determine the target instance size, and *sizes* informs the sizing strategy about previous deviations from the target number of weights.
15	Fits the model instance to the training data.
16	Submits the test data to the now trained model instance.
17	Evaluates the performance of the model instance across the metrics that are appropriate for the current *task*, and stores the results in the list L. After the loop started in row 12 completes, this list will contain *ss* items.
18	Determines the actual size of the current model instance and updates the history in *sizes*.
20*	Uses the sample of performance data in L to compute a bootstrapped confidence interval for each metric that was assessed. The resulting statistics are collected in a hash table and stored in a cell of the *results* tensor, whose dimensions are indexed by the triple (dataset, model, config). After the loop started in row 3 completes, *results* holds the evaluations for Polygrid across all datasets and configs, and the evaluations of the alternative models for the configs identified in the first stage of the evaluation.
24	Returns the *results* tensor updated with the performance data of the alternative models.

Source: The author.

Note: The rows in which Algorithms 10 and 11 mostly diverge are marked with an asterisk.

5.1.6 Data analysis

To support our first goal, which requires the comparison of multiple models based on their performance over multiple datasets, we adopted a popular method in machine learning research, advanced by Demšar (2006) and Benavoli, Corani and Mangili (2016). The method consists of applying the Friedman test to the data to test for significant differences in performance among models. If a significant difference is found, then pairwise comparisons of average ranks using the Wilcoxon signed-rank test are performed. Since multiple comparisons inflate the family-wise error rate, the significance level α is adjusted accordingly using the Bonferroni's or Holm's method. Luckily, this method has been automated by software such as the critdd package. However, every statistical method comes with assumptions, and a critical one is not met in our case: the datasets must provide independent observations. The issue is that the three original datasets were augmented with synthetic assignments to build our collection of 15 datasets which, thus, are not independent. In fact, we found that the method implemented in critdd was not able to discern the accuracy of the Random model from that of the remaining ones, which we take as evidence that the violation of the independence assumption is too severe.

To overcome this issue, we shift from a significance testing approach to a more deterministic one. Pairwise comparisons of average ranks are still the core of the method but signed rank tests are replaced by dominance tests. The latter are based on a matrix A which counts the number of times a model dominates another on a given metric. For example, assume j_0 refers to the Polygrid model, and j_1 to the Linear model. The value of $A_{j_0 j_1}$ is increased by one whenever the confidence intervals for accuracy do not overlap, and the point estimate for Polygrid is greater than that of the Linear model. Finally, to rank the models according to some metric, we combine the average ranks with the dominance data using a method that is based on the following analogy[10]:

- The top-ranking model (in terms of average rank) is "hired" to lead the first echelon.
- Once hired, the leader must staff the first echelon with competitive models. A model j_1 is competitive with the leading model j_0 if $abs(A_{j_0 j_1} - A_{j_1 j_0}) \leq 1$. The value $A_{j_0 j_1}$ is the number of times that model j_0 dominates the performance of model j_1. The upper bound of the inequality accounts for settings with an odd number of datasets.
- Once all models that are competitive with the leader have been hired, the next free top-ranking model is hired to lead the second echelon, and the process restarts.
- The process ends when all models have been hired.

As a result, the ranking of the models is represented by a hierarchy of echelons, with each echelon being represented by a non-empty set of model indices. Figure 24 shows the ranking produced by this method. The following statements are true about the hierarchies shown in the figure: the leading model in the first echelon dominates the leading model in the second echelon, which dominates the leading model in the third echelon, and so on. Moreover, any model in an echelon is competitive with the leading model in that same echelon[11]. Note that this graphical representation conveys ranking information that is structurally equivalent to that found in critical difference diagrams, which are used in the standard method mentioned earlier (Demšar, 2006), or in Hasse graphs, which have recently been proposed in emerging evaluation methodologies (Jansen *et al.*, 2024).

Regarding the remaining goals of this evaluation, they will be addressed ad hoc, since there are no standard approaches in the literature. We anticipate that the second goal will be mostly based on the comparison of computational graphs of Polygrid and the alternative models, and the third goal will be based on the analysis of the search space.

[10] We explore the analogy "models are like persons" to make the explanation more intuitive.

[11] It is an open matter whether the following statements are true: (a) any model in the first echelon dominates the leading model of the second echelon, and (b) any model in the first echelon dominates all models in the second echelon, (c) this approach avoids the issue pointed out by Benavoli, Corani and Mangili (2016), regarding the inversions in established rankings induced by the introduction of new models in the evaluation.

5.2 Results

The results of the first stage of the evaluation are shown in Tables 16 and 17. The indices of the configs in which Polygrid achieves the highest performance on some dataset and metric are listed on Table 16. Of the 3,456 Polygrid configurations generated by Algorithm 10, only 47 instances were selected for some dataset and metric. Many configs appear more than once. The specification of a config can be consulted by looking for its index in Table 17. For example, the config in which Polygrid achieved the highest accuracy on the whoqol dataset has index 166, and corresponds to ($nspd = 1$, $na = 2$, *vorder: averages*, *annulus type: r-invariant*, *sector type: cover*, *solver: ridge*, *cutoff scheme: single*).

The results of the second stage of the evaluation are illustrated in Figures 21, 22, and 23. In the first figure, the results achieved by Polygrid and the alternative models on the multiclass datasets are shown side by side. The second and third figures show similar results for the multilabel and label-ranking datasets, respectively. A first inspection at the detailed results shows that Polygrid was competitive with the alternative models in multiclass and multilabel datasets, but was less successful in label ranking datasets. This conclusion is supported by the statistical analysis summarised in Figure 24, in which the model rankings for each metric are displayed in a diagrammatic form, sided by the resulting dominance matrix used to group the models into distinct levels of performance.

Table 16 – Indices of the best Polygrid configs per metric and dataset (Stage 1)

	multiclass and multilabel datasets								
	whoqol			ampiab			elsio1		
metric	base	ml-11	ml-22	base	ml-11	ml-22	base	ml-11	ml-22
accuracy	166	3168	3304	1952	2806	3193	2995	3429	2275
hammingl	166	3168	3434	1952	166	3185	2995	3267	2115
f1.micro	166	3424	3450	1952	1494	3185	2995	3267	2275
f1.macro	166	3349	3402	1952	1164	2993	3283	3267	2275
f1.weigh	166	3349	3450	1952	2313	3185	3285	3429	2275
jaccsim	6	3349	1830	42	1884	1507	121	2976	2646

	label ranking datasets								
	whoqol			ampiab			elsio1		
metric	base	lr-11	lr-22	base	lr-11	lr-22	base	lr-11	lr-22
ktau	-	3032	3081	-	3185	1457	-	2799	3281
lracc	-	3394	2757	-	3345	2873	-	3413	2173
lrloss	-	3394	3317	-	3184	2769	-	3408	2131

Source: The author.

Note: Results from the first stage of the offline evaluation. The upper block shows the best configs for multiclass and multilabel datasets and the lower block for label ranking datasets. The details of each config can be recovered by consulting its corresponding index in Table 17.

Table 17 – Descriptions of the best Polygrid configs found in stage 1, by index

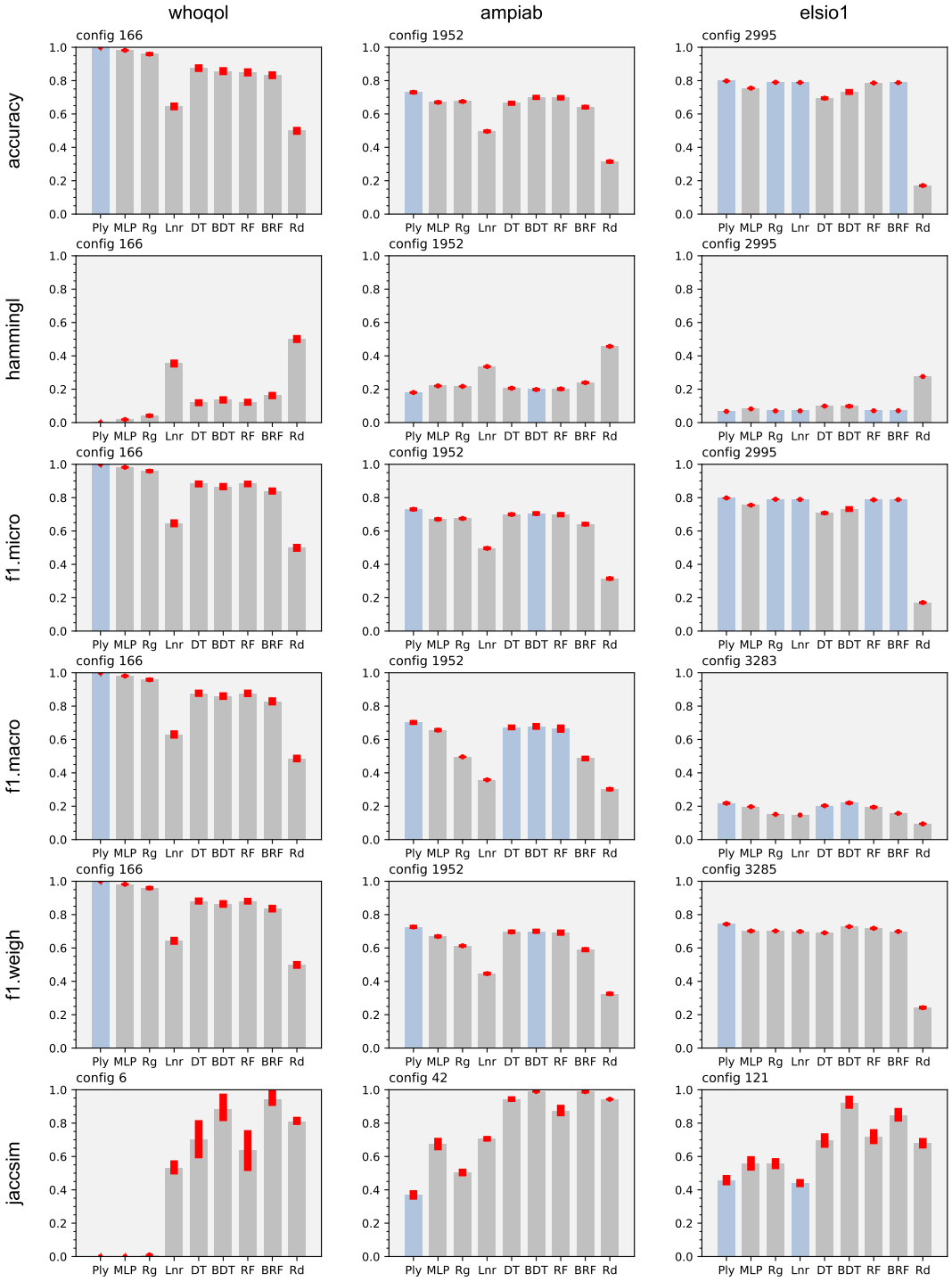
index	description	index	description
6	(1, 1, avg, s-invt, cover, ridge, single)	2995	(3, 5, msr, r-invt, cover, lstsqsym, multiple)
42	(1, 1, avg, tree, miss, lstsqsym, single)	3032	(3, 6, avg, s-invt, miss, lstsq, single)
121	(1, 1, msr, r-invt, miss, lstsq, multiple)	3081	(3, 6, rho, s-invt, miss, lstsq, multiple)
166	(1, 2, avg, r-invt, cover, ridge, single)	3168	(3, 7, avg, s-invt, cover, lstsq, single)
1164	(2, 1, avg, s-invt, miss, lstsquni, single)	3184	(3, 7, avg, r-invt, cover, lstsq, single)
1457	(2, 3, avg, r-invt, cover, lstsq, multiple)	3185	(3, 7, avg, r-invt, cover, lstsq, multiple)
1494	(2, 3, rho, s-invt, cover, ridge, single)	3193	(3, 7, avg, r-invt, miss, lstsq, multiple)
1507	(2, 3, rho, r-invt, cover, lstsqsym, multiple)	3267	(3, 7, msr, s-invt, cover, lstsqsym, multiple)
1830	(2, 5, msr, s-invt, cover, ridge, single)	3281	(3, 7, msr, r-invt, cover, lstsq, multiple)
1884	(2, 6, avg, s-invt, miss, lstsquni, single)	3283	(3, 7, msr, r-invt, cover, lstsqsym, multiple)
1952	(2, 6, rho, tree, cover, lstsq, single)	3285	(3, 7, msr, r-invt, cover, lstsquni, multiple)
2115	(2, 7, msr, s-invt, cover, lstsqsym, multiple)	3304	(3, 7, msr, tree, miss, lstsq, single)
2131	(2, 7, msr, r-invt, cover, lstsqsym, multiple)	3317	(3, 8, avg, s-invt, cover, lstsquni, multiple)
2173	(2, 8, avg, s-invt, miss, lstsquni, multiple)	3345	(3, 8, avg, tree, cover, lstsq, multiple)
2275	(2, 8, msr, r-invt, cover, lstsqsym, multiple)	3349	(3, 8, avg, tree, cover, lstsquni, multiple)
2313	(3, 1, avg, s-invt, miss, lstsq, multiple)	3394	(3, 8, rho, tree, cover, lstsqsym, single)
2646	(3, 3, rho, s-invt, cover, ridge, single)	3402	(3, 8, rho, tree, miss, lstsqsym, single)
2757	(3, 4, avg, r-invt, cover, lstsquni, multiple)	3408	(3, 8, msr, s-invt, cover, lstsq, single)
2769	(3, 4, avg, tree, cover, lstsq, multiple)	3413	(3, 8, msr, s-invt, cover, lstsquni, multiple)
2799	(3, 4, rho, s-invt, miss, ridge, multiple)	3424	(3, 8, msr, r-invt, cover, lstsq, single)
2806	(3, 4, rho, r-invt, cover, ridge, single)	3429	(3, 8, msr, r-invt, cover, lstsquni, multiple)
2873	(3, 4, msr, tree, miss, lstsq, multiple)	3434	(3, 8, msr, r-invt, miss, lstsqsym, single)
2976	(3, 5, msr, s-invt, cover, lstsq, single)	3450	(3, 8, msr, tree, miss, lstsqsym, single)
2993	(3, 5, msr, r-invt, cover, lstsq, multiple)	—	

Source: The author.

Legend: This is the order in which hyperparameters appear in a tuple: (number of sectors per domain, number of annuli, ordering of the vertices, type of annulus, type of sector, type of solver, cutoff scheme). Some values have been abbreviated: averages (avg), measures (msr), s-invariant (s-invt), and r-invariant (r-invt).

Before we discuss how the goals of the evaluation are supported by these results, we want to address two surprising patterns that appear in the dominance matrices. In principle, the performance of the Random model should not dominate any of the alternative models. This implies that the last row of any dominance matrix in Figure 24 should be filled with zeroes. Moreover, if we extend this principle to state that the performance of any of the models should dominate the Random model, then the last column of these dominance matrices should be filled with 9 (in multilabel metrics) or 6 (in label ranking metrics), except in the last row. These two patterns are observed for the accuracy, the Hamming loss, and the micro-averaged F1 score, but they break for the remaining metrics. The reasons for this vary. For example, in the macro-averaged F1 score, the Random model

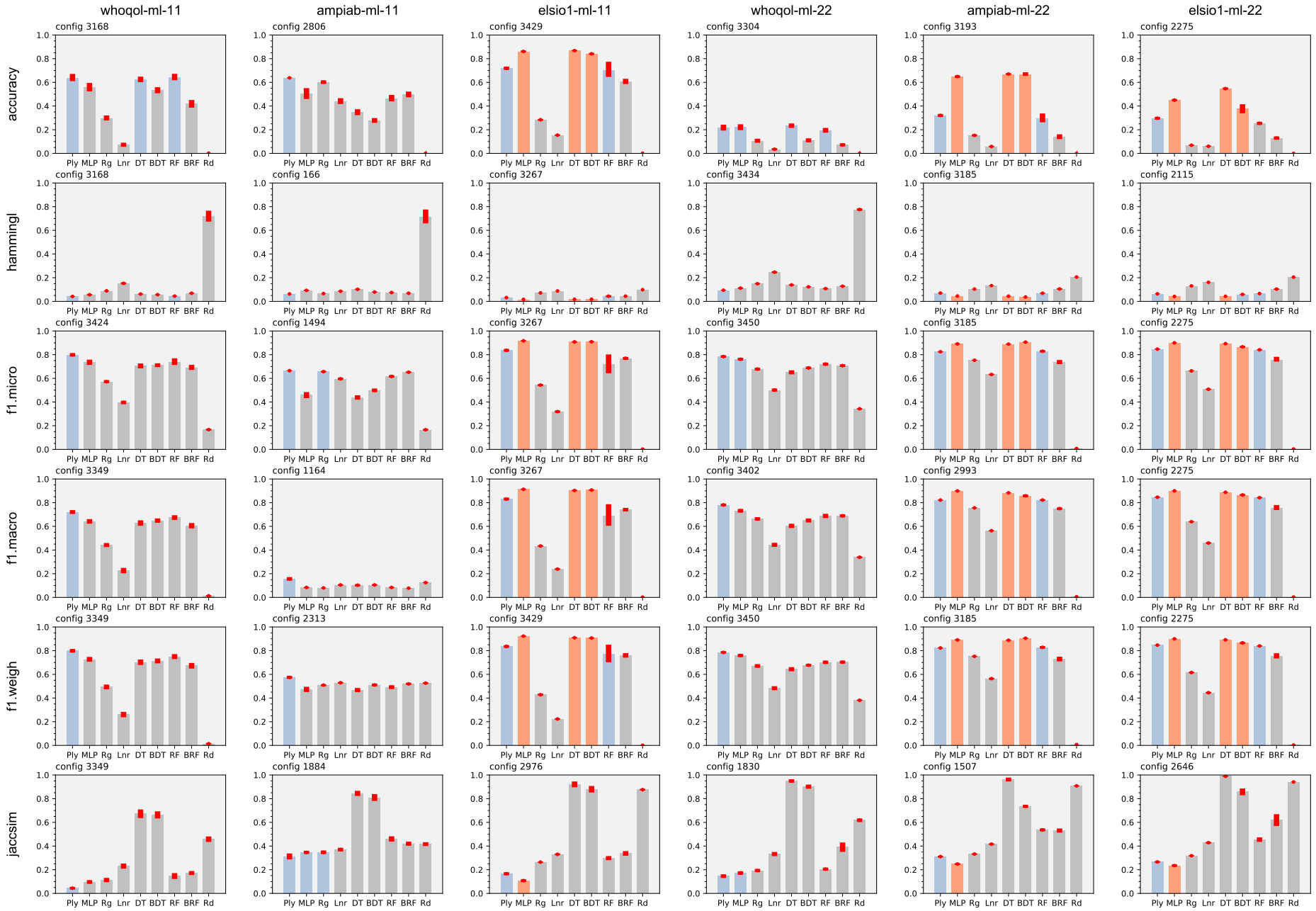
Figure 21 – Results from the offline evaluation on the multiclass datasets



Source: The author.

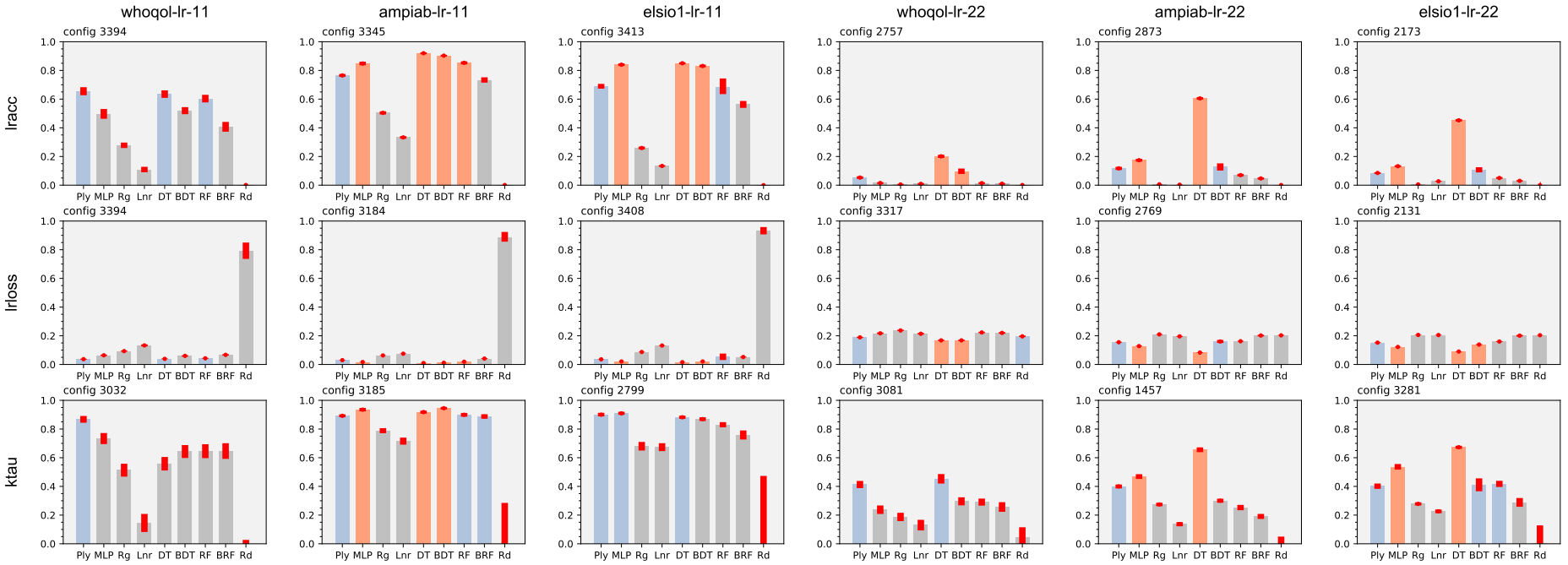
Legend: Each chart shows a comparison between Polygrid and the alternative models regarding their performance on some dataset and with respect to some metric. The position of a chart encodes its relevant dataset and metric by the column and row it occupies in the grid, respectively. The config is identified at the top left of each chart. The names of some models have been abbreviated: Polygrid (Ply), Linear (Lnr), Ridge (Rg), BRDT (BDT), BRRF (BRF), and Random (Rd). The red rectangle at the top of a bar represents the confidence interval for that measurement. A blue bar indicates that its confidence interval overlaps with Polygrid's. A grey bar means lower performance than Polygrid, and an orange bar means higher performance. A worksheet with the results of the evaluation is available as supplementary material.

Figure 22 – Results from the offline evaluation on the multilabel datasets



Source: The author.

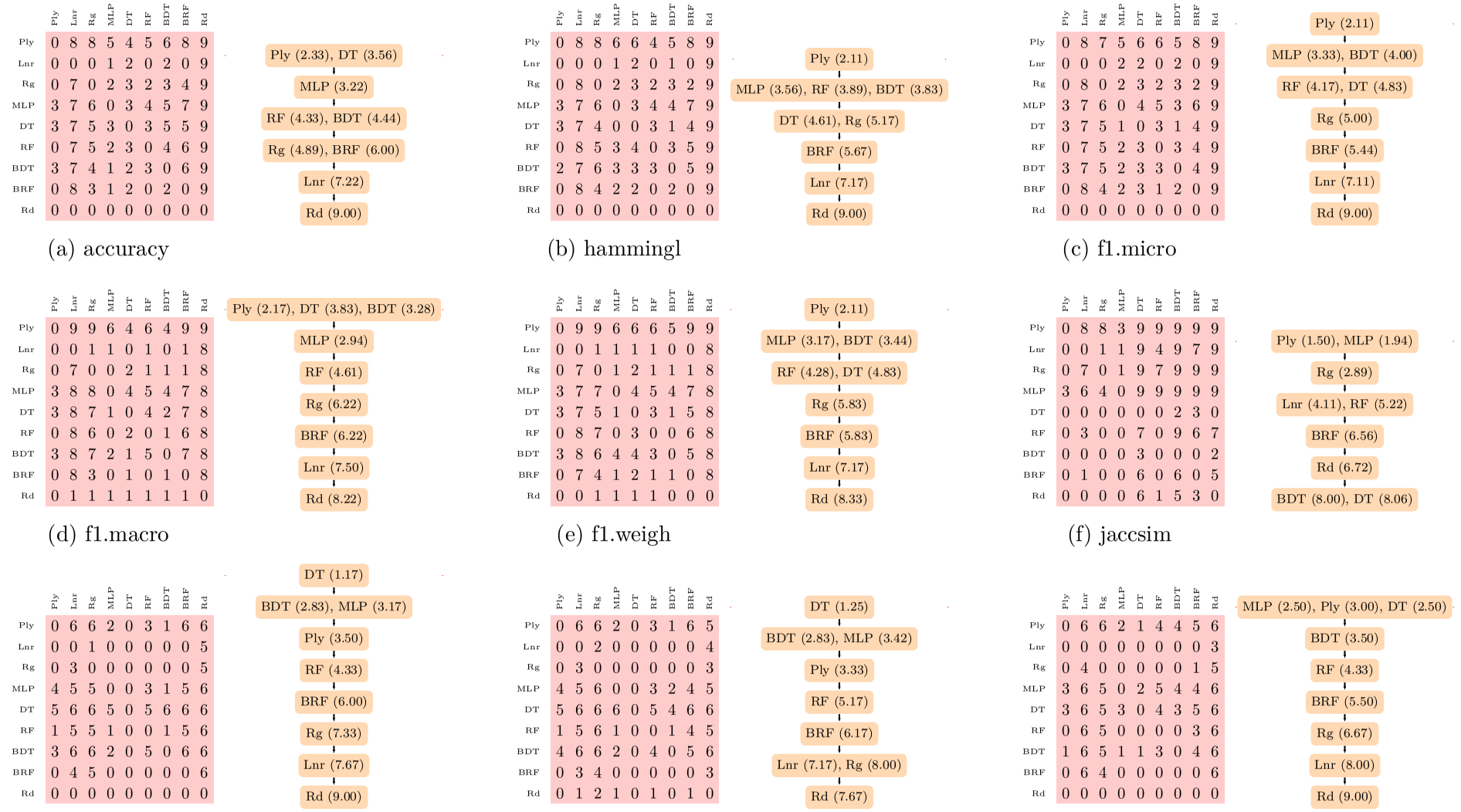
Figure 23 – Results from the offline evaluation on the label ranking datasets



Source: The author.

Legend: Each chart shows a comparison between Polygrid and each of the alternative models regarding their performance on some dataset and with respect to some metric. The row and the column in which a chart appears in the grid indicate its relevant metric and dataset, respectively. The relevant config is informed at the top left of each chart. The names of some models have been abbreviated: Polygrid (Ply), Linear (Lnr), Ridge (Rg), BRDT (BDT), BRRF (BRF), and Random (Rd). The red rectangle at the top of a bar represents the confidence interval for that measurement. A blue bar indicates that its confidence interval overlaps with Polygrid's. A grey bar means lower performance than Polygrid, and a red bar means higher performance. In both grey and red cases, the confidence interval do not overlap with Polygrid's. A worksheet with the results of the evaluation is available as supplementary material.

Figure 24 – Comparative results from the offline evaluation by each metric



Legend: Models ranked by their performance in each metric. The matrix in red corresponds to the dominance matrix A discussed in the text. An element $A_{j_0 j_1} = c$ means that model j_0 dominated the performance of model j_1 in c datasets, with respect to some metric. The diagram represents the ranking of the models, also discussed in the text. The number that appears between parenthesis indicates the average rank obtained by the model. As before, the names of some models have been abbreviated: Polygrid (Ply), Linear (Lnr), Ridge (Rg), BRDT (BDT), BRRF (BRF), and Random (Rd).

dominates the performance of the alternative models in the ampiab-ml-11 dataset (see config 1164 in Figure 22). This is due to the severe imbalance of the dataset, with most instances assigned to the label "Specialty" (86 out of 128), and most labels assigned to less than five instances each (AME, Social, NCI, CER, Exams+, and Other internal and external procedures). For these low-frequency labels, the alternative models are prone to obtain null precision, while the Random model obtains small precision and recall scores. The class imbalance also explains the last row/column pattern for the weighted F1 score.

Regarding the lrloss metric, the Random model dominates the performance of several alternative models on the whoqol-lr-22 dataset, and dominates the BRRF model on the ampiab-lr-22 dataset. This pattern is not observed in the *-lr-11 datasets. For the even-ending configs in Figure 23 (3394, 3184, and 3408), the lrloss of the Random model is much higher because these configs impose a single cutoff to be shared by all labels. In this circumstance, the Random model is prone to set the cutoff within the [0.0, 0.5] interval, which increases the cardinality of the predicted subset to a level much higher than the baseline of 1.09 of these datasets. If multiple cutoffs are allowed, the lrloss of the Random model decreases substantially, but still not enough to dominate other models, except for the Linear model on the whoqol-lr-11 dataset. This is possibly an artefact of the way the lrloss metric interacts with the choice of representation we adopted for label ranking datasets. Since rankings may vary in size, we use "-1" to encode a filler in incomplete rankings (see Section 4.1). The lrloss takes these fillers into account when computing mismatches, possibly giving the Random model an undue advantage.

Finally, the last row/column pattern breaks for the Jaccard similarity metric for a different reason: all tree-based models (DT, BRDT, RF, and BRRF) produce extreme scores. For example, Figure 25a shows a scales diagram for the Linear model on the whoqol dataset. The scores for the true positive cases of "Poor QOL" in the test sample are scattered throughout the interval $tp = [0.23, 0.91]$, while the scores for the true negative cases are in $tn = [0.36, 1.06]$. The Jaccard similarity score of this setting is $\frac{\mu(tp \cap tn)}{\mu(tp \cup tn)} = 0.66$. Figure 25b shows the same diagram for the DT model. True positive cases of "Poor QOL" in the test sample are confined to the interval $tp = [0.11, 1.0]$, while the true negative cases are in $tn = [0.0, 1.0]$. The Jaccard similarity score is $\frac{\mu(tp \cap tn)}{\mu(tn \cup tn)} = 0.89$.

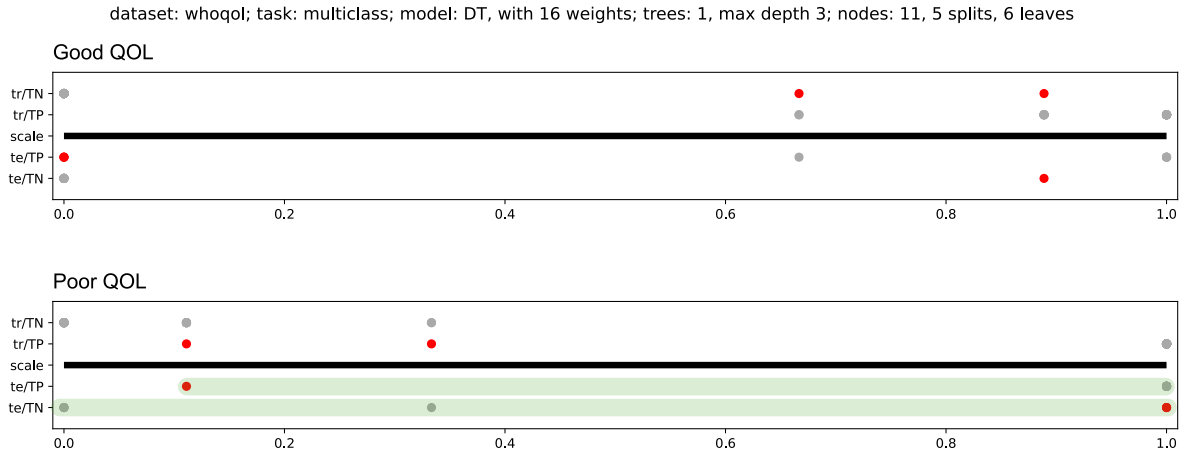
5.3 Discussion

As said before, the statistical analysis summarised in Figure 24 shows that Polygrid is competitive with the alternative models in multiclass and multilabel datasets, but was less successful in label ranking datasets. In this section, we explore these results to approach the goals of the evaluation. We start with the first two goals, which aim to place Polygrid in the rankings of alternative models considering the tasks, datasets, and metrics that described in Section 5.1, as well as identifying situations in which an alternative

Figure 25 – Scale diagrams for the WHOQOL dataset, config 166

(a) Diagram for the Linear model

(b) Diagram for the DT model



Source: The author.

Legend: The *tp* and *tn* intervals for "Poor QOL" in the test sample are highlighted in green. In tree-based regression models, the number of distinct prediction values correspond to the number of leaves of the tree. Some leaves are bound to predict the extreme of the scale, and this implies that *tp* and *tn* are bound to include one of (or both) the extremes.

model achieves a performance that is markedly superior to that obtained by the Polygrid model in an equivalent setting. We remind the reader that the following analyses are made in the spirit of learning how Polygrid can be improved to operate in the specific domain of application we described as gerontological primary care, rather than presenting Polygrid as a general alternative to models that are traditionally employed to solve multilabel and label ranking tasks. The discussion is organised by type of task and evaluation goals. We discuss how Polygrid performed when compared to the alternative models and analyse the reasons why the alternative models attained better results. The last section shares some lessons we learned about how to select the hyperparameters for the Polygrid model.

5.3.1 The performance of Polygrid on multiclass datasets

As summarised in Figure 21, the performance of the Polygrid model has dominated the alternative models in multiclass datasets, with a few situations in which some alternative model shows comparable performance across multiple metrics. For example, BRDT ties with Polygrid on the ampiab dataset across all metrics except (subset) accuracy and jaccsim, and the Linear and Ridge models tie with Polygrid on the elsio1 dataset on several metrics. So, we are interested in clarifying why Polygrid performed so well.

Recall from Section 5.1.1 that a common application of standardised instruments in gerontological studies regards their use to stratify the target population with respect to some risk. This is why the multiclass datasets were included in our collection. The synthetic assignments created for these datasets reflect decision boundaries based on sum-scores. For example, the decision boundary advanced by Silva *et al.* (2014) for the whoqol dataset is defined as $\sum_j \mathring{x}_{ij} \geq 60$. Similar statements can be made to the ampiab dataset (Andrade *et al.*, 2020) and, to some extent, to the elsio1 dataset (Aliberti *et al.*, 2022).

Our hypothesis is that Polygrid dominated the alternative models because it is better suited to express decision boundaries with such a structure. This is due to the monotonic relationship between sum-scores and area-scores[12], which becomes stronger inasmuch as the requirements for psychometric data described in Section 4.1 are satisfied. This relationship is stronger in the whoqol and elsio1 datasets than in the ampiab dataset.

We present the following evidence in support of this hypothesis. Polygrid achieves its top accuracy on the whoqol dataset with config 166, which specifies a small model ($nspd = 1, na = 2,$ with 18 weights) based on the ridge solver. In contrast, its top accuracy on the elsio1 dataset, $0.798 \in [0.792, 0.803]$, is achieved by a much larger model specified by config 2995 ($nspd = 3, na = 5,$ with 450 weights), which is based on the lstsqsym solver[13]. However, a closer look at the results reveals that a small model based on the ridge solver shows a comparable performance: config 46 ($nspd = 1, na = 1,$ with 36 weights) achieves $0.794 \in [0.79, 0.799]$. Recall that the ridge solver includes an intercept parameter, which may play the role of a constant in the decision boundaries such as $\sum_j \mathring{x}_{ij} \geq C$.

In agreement with our hypothesis, a similar analysis fails for the ampiab dataset, on which Polygrid achieves its top accuracy of $0.73 \in [0.718, 0.742]$ with config 1952, ($nspd = 2, na = 6,$ with 180 weights) based on the lstsq solver. The highest accuracy achieved with the ridge solver is $0.697 \in [0.687, 0.706]$ (config 3239, $nspd = 3, na = 7,$ with 318 weights). This is a considerably larger model ($318 > 180$ weights) with a lower performance.

[12] The area-score corresponds to the area of an assessment polygon. The relationship between sum- and area-scores is developed in Appendix A, which also presents evidence in support of the relationship.

[13] The lstsqsym solver is identical to the lstsq solver, except for a change in how assignments are encoded: $y_{ij} \in \{-1, 1\}$ instead of $y_{ij} \in \{0, 1\}$. There is evidence that this decision may be consequential for regression-like models for multilabel classification (Jia; Liu; Zhang, 2024).

5.3.2 The performance of Polygrid on multilabel datasets

As shown in Figure 22, the Polygrid model obtained mixed results on multilabel datasets. More specifically, Polygrid has dominated the alternative models on both the whoqol datasets (whoqol-ml-11 and whoqol-ml-22) and on the ampiab-ml-11 dataset. However, Polygrid was consistently dominated by MLP, DT, and BRDT on both elsio1 datasets (elsio1-ml-11 and elsio1-ml-22) and on the ampiab-ml-22 dataset. So, we are now interested in clarifying what the alternative models did differently from Polygrid to grant them higher performance in these settings. As said before, the ultimate goal is to identify ideas to improve the performance of the Polygrid model to be explored in future work.

As shown in Figure 26a, the architecture of the MLP model used in this evaluation has a single hidden layer with sigmoid activation. The number of neurons in the hidden layer, $|h|$, is calculated so that the size of the MLP instance matches that of the competing Polygrid instance. The output of a neuron in the hidden layer is the inner product between the input vector X_i and a weight vector associated with that neuron, squashed by the activation function. This means that the hidden layer projects the input vector X_i into a feature space embedded in $\mathbb{R}^{|h|}$. Regarding the output layer, it has n neurons (one for each label), and each represents an inner product between a feature vector in $\mathbb{R}^{|h|}$ and a weight vector (without activation), generating an output space embedded in $\mathbb{R}^n$.

A forward computational graph of the Polygrid model is illustrated in Figure 26b. Its output layer is structurally similar to that of the MLP model in that it performs an inner product between a feature vector (i.e., the output of the "annular sectors" layer) and the weights of the output layer. However, the models have different feature extraction methods. Although the MLP's hidden layer and the Polygrid's "annular sectors" layer are both bounded[14], they have different dimensionalities and structures. For example, in Figure 26, while both instances have 16 weights, the MLP instance extracts feature vectors with $|h| = 2$ dimensions, whereas Polygrid extracts vectors with $n_{as} = 8$ dimensions. Moreover, MLP uses learned weights for feature extraction, meaning that the weights are optimised to hit the targets during the training. In contrast, Polygrid's feature extraction method depends solely on the input vector X_i. To test the first hypothesis, that the observed difference in performance is due to the difference in the dimension of the feature spaces, we compared Polygrid's performance in feature spaces with nearly the same dimensionality:

- In the elsio1-ml-11 dataset, Polygrid's highest accuracy was achieved with config 3429. The competing MLP instance had 77 neurons in the hidden layer. A Polygrid instance with config 2979 ($n_s/d = 3, n_a = 5$) extracts feature vectors with 74 dimensions, and with config 2993 ($n_s/d = 2, n_a = 8$) extracts vectors with 80 dimensions. The accuracy of these instances is much lower than that of the competing MLP instance.

[14] For the MLP model, $h_t \in (0, 1)$, and for the Polygrid model, $s_r \in (0, 1/2]$.

- In the elsio1-ml-22 dataset, Polygrid's highest accuracy was achieved with config 2275. The competing MLP instance had 63 neurons in the hidden layer. A Polygrid instance with config 1995 ($n_s/d = 2, n_a = 6$) extracts feature vectors with 60 dimensions, as does an instance with config 2835 ($n_s/d = 3, n_a = 4$). Again, the accuracy of these instances is much lower than that of the competing MLP instance.

- In the ampiab-ml-22 dataset, Polygrid with config 3193 achieves the highest accuracy. The competing MLP instance had 81 neurons. A Polygrid instance with config 2201 ($n_s/d = 2, n_a = 8$) extracts feature vectors with 80 dimensions. The accuracy of this instance is much lower than that of the competing MLP instance.

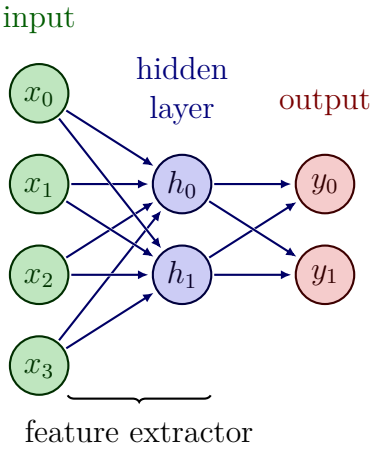
Considering these results, we dismiss the first hypothesis. The second hypothesis, that the difference in performance should be ascribed to the structure of MLP's feature extraction (inner products with learned weights), remains. If this is the case, an adaptive version of the Polygrid's feature extractor is promising. By adaptive we refer to the ability to use both assessment and assignment data to learn optimal boundaries of the annuli (or the sectors). This corresponds to designing an alternative to Algorithm 3. However, the argument for Polygrid's interpretability relies on the way in which the explanation diagram visually encodes the forward computational graph and the model‘s architecture preserves the meaning of the input data, so any adaptations should consider these aspects.

The analysis of the DT model is also promising, as we shall see. It obtained the highest accuracy in the three datasets in which Polygrid was dominated by some alternative model. Basically, the DT model learns how to partition the input space into hypercubes and assign each to a target label. For this purpose, the model uses data from both the input data and the target labels. The boundaries of these hypercubes are specified by multiple "thresholds" defined along each dimension of the input space.

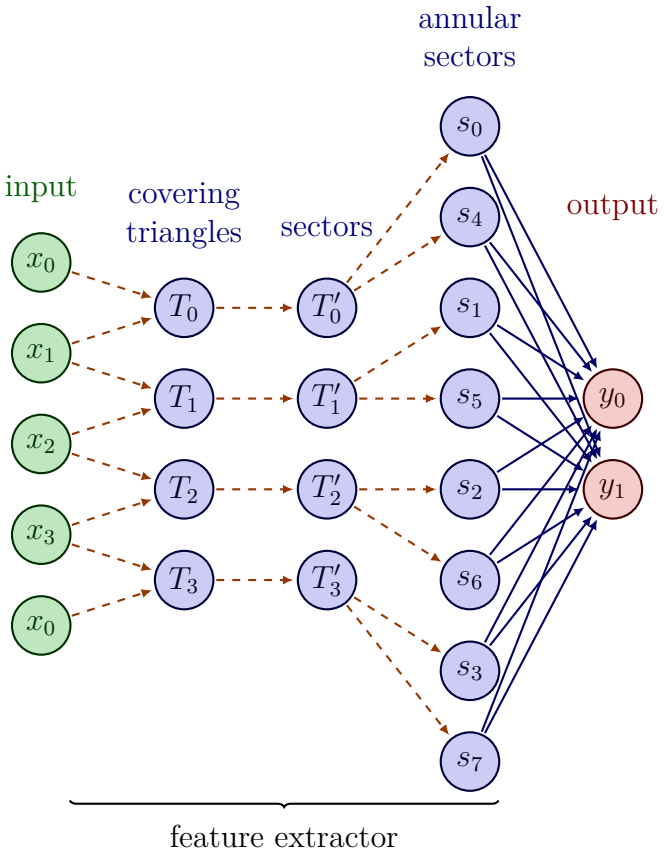
This characteristic is interesting because it resembles the way Polygrid's partitions the input space. The boundaries learned by a decision tree can be used to specify the annuli of a Polygrid instance, as shown in Figure 18. In that example, only the top $(n_a - 1)$ boundaries were used. We explored this idea further with the elsio1-ml-11 dataset. A Polygrid instance with config 3429 achieved the highest accuracy in this dataset. Its competing DT instance learned 43 distinct boundaries. The average distance between two such boundaries was about 0.02, and the minimum distance was about 0.0002. The average distance would require a Polygrid instance with 50 annuli, and the minimum distance implies an instance whose diagram cannot be inspected by the user. We rounded the boundaries to two decimals and constrained the minimal difference to 0.02, resulting in about 24 annuli. The results showed that an instance with these boundaries attained modest improvement when compared to the Polygrid for config 3429, and was still dominated by DT. A similar analysis applied to the elsio1-ml-22 and ampiab-ml-22 datasets lead to

Figure 26 – Examples of forward computational graphs for MLP and Polygrid

(a) MLP, 2 neurons in the hidden layer, with 16 weights



(b) Polygrid, $n_s/d = 1$, $n_a = 2$, sector type "miss", solver "lstsq", with 16 weights



Source: The author.

Legend: Forward computational graphs of MLP and Polygrid instances fit to the whoqol dataset, with $d = 4$ inputs, and $n = 2$ outputs. The blue solid edges represent learned weights, and the red dashed edges signal message passing. All weights in an MLP instance are learned, including those used for feature extraction. In contrast, Polygrid's learned weights are confined to the output layer, since its feature extraction method is algorithmically fixed. Note that MLP neurons have an extra weight for bias, which are not represented above.

similar results. Our conclusion is that this idea is promising, but the challenge is to learn to partition the input space (i.e., to design an alternative to Algorithm 3) that combines both input and target data with no sacrifice to the Polygrid's interpretability.

Finally, the third model that dominated Polygrid's performance on the three datasets is the BRDT model, which fits a DT model for each label separately. At first sight, since each DT would possibly partition the input space differently, one may consider two paths to adapt ideas from this model: (a) to allow for each label to have its own partitioning, or (b) to create a single partition that combines all separate partitions. The second one seems more promising, since the first alternative would disrupt the interaction of the user with the Polygrid's diagram. However, it must be said that the DT model, which fits a single decision tree, is simpler and achieved higher accuracy. Thus, it seems reasonable to prioritise the ideas for improvement coming from the DT model.

As a closing remark, we want to address a curious pattern in the results shown in Figure 22 regarding the Hamming loss for the Random model. The loss is much higher for even-ending configs than for odd-ending configs. This happens because even configs impose a single cutoff to be shared by all scales, whereas odd configs allow for each label to have a locally optimal threshold. The algorithm that optimises the thresholds seeks to minimise the occurrence of false negatives. In this setting, the Random model with a single threshold is prone to set the cutoff within the [0.0, 0.5] interval, generating a large number of false positives. The use of multiple thresholds reduces the Hamming loss, but does not improve the subset accuracy, as one would expect.

5.3.3 The performance of Polygrid on label ranking datasets

As shown in Figure 23, the Polygrid model was predominantly dominated on label ranking datasets. Exceptions are the whoqol-lr-11 dataset, in which Polygrid is dominant, and elsio1-lr-11 and whoqol-lr-22 datasets on the Kendall's tau metric. The alternative models that dominated most scenarios are the same as in the multilabel classification: MLP, DT and BRDT, with the DT model standing out on datasets with 22 labels. Consequently, all recommendations for improving Polygrid to perform multilabel classification would also benefit this task. However, Polygrid applied to label ranking tasks suffers from pains that are not addressed by those improvements. For this reason, this section focuses on explaining the origin of these new pains and point out some ideas to tackle them. We start by justifying why this analysis was made on the whoqol datasets with 22 labels, and then inspect the results to make the novel issues explicit, and develop some ideas.

It must be noted that all datasets with 11 labels have a cardinality around 1.08, as indicated in Table 12. In this setting, the expected difference in performance of any model solving a multilabel classification task and a label ranking task should be minimal. This is because both tasks could be closely described as a multiclass classification task: the

challenge consists of assigning a single label to each case in the dataset for the majority of cases. The results shown in Figures 22 and 23 confirm this expectation, with the exception of the ampiab-ml-11 and ampiab-lr-11 datasets. The explanation for this exception lies in two differences between these datasets, which do not occur in the other datasets. First, in the ampiab-ml-11 dataset, the targets are real patient referral data, while the targets are synthetic in the ampiab-lr-11 dataset, as described in Section 5.1.1. Second, the 128 cases in the ampiab-ml-11 dataset are a sample of the 510 cases in the ampiab-lr-11 dataset.

Regarding the datasets with 22 labels, they all have cardinality 4.54, meaning each case in a dataset is assigned to 4 or 5 labels on average. In this setting, the expected difference in performance of any model solving a multilabel classification task and a label ranking task must be significant. This is because label ranking tasks allow for an additional dimension of error: not only can false negatives and positives occur, but also the order in which the positive labels are presented can be wrong. The results shown in Figures 22 and 23 confirm this expectation. They also show an interesting contrast involving the whoqol datasets with 22 labels: Polygrid dominated the alternative models on the multilabel classification dataset, but is dominated in the label ranking dataset. We will use this contrast to investigate the drop in Polygrid's performance between tasks.

The results obtained from a Polygrid instance trained on the whoqol-lr-22 dataset with config 2757 are shown in Figure 27. The targets in the test partition during the evaluation are reproduced in the "Targets" matrix, and the "Predicted" matrix reproduces the output generated for the corresponding cases. The rightmost matrix, in which errors are classified into distinct categories and counted, allows us to develop a first conclusion:

- Of the 20 test cases, the prediction matches the target in only one;
- Among the 19 cases with some kind of error, 8 cases show errors related to false negatives and false positives, but no errors involving the order of labels;
- Of the remaining 11 cases, all involve some error in the order in which the labels are presented. One of these cases does not involve false negatives or false positives.

This profile suggests that even if we could eliminate the occurrence of false negatives and false positives, hopefully by pursuing the recommendations made in the preceding section, this would clear about half of the cases in this example. However, the situation seems to be more complicated: compared to the results in Figure 28, obtained by a Polygrid instance with the same config, trained on the corresponding multilabel dataset (whoqol-ml-22) with the same train/test split, a huge increase in false negatives is seen: from 7 (multilabel) to 22 (label ranking). This increase is accompanied by a moderate to large decrease in false positives: from 28 (multilabel) to 20 (label ranking). Based on this comparison, our second

conclusion is that, when moving from a multilabel classification to a label ranking setting, the Polygrid model may have "lost" some of its ability to assign proper labels to cases.

We believe that this increase in error may be due to (a) some difference between Algorithms 5 and 8, used by Polygrid to learn weights for multilabel and label ranking tasks, respectively; and (b) the method we use to convert the matrix Y with targets for label ranking into the matrix U that is actually fed to the algorithms just mentioned. Thus, our recommendation is to focus on improved versions of these methods.

5.3.4 Is the evaluation design biased towards Polygrid?

The unusual design adopted in this evaluation may raise questions about its fairness. After all, the standard evaluation in works that introduce new models into mature domains does not impose restrictions on the size of the models being compared (Demšar, 2006). However, as justified in the beginning of this chapter, our interest in promoting Polygrid's interpretability and performance imposes the need to control for model size[15].

For this reason, we want to address the argument that the method that controls the size of alternative models during evaluation, described in Section 5.1.5, introduces a systematic bias in favour of the Polygrid model or to the detriment of alternative models. More precisely, we want to address the argument that this method amplifies the variance of model size (as a random variable), and this results in a reduced accuracy reading.

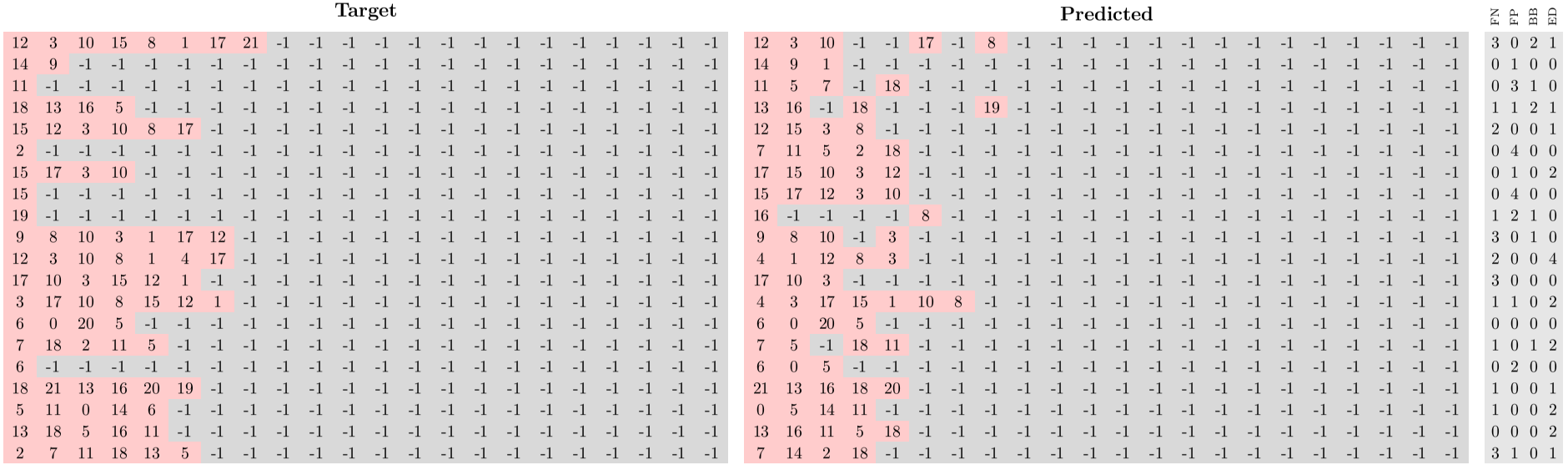
In some settings, controlling for model size is impossible because the size of some models is fixed by the dimensionality of the data (i.e., the number of features and targets). This is the case of the Linear and Ridge models, which take up less than 150 weights for the datasets and configs in Table 16. For the same settings, Polygrid takes up about 1,164 weights on average. Since the mechanism we use to control for model size does not exert an effect on the size of these models, no variance that could bias the performance reading is introduced. Thus, the argument about bias must be dismissed in these cases.

In contrast, the MLP model is on the other extreme of the control spectrum, as the size of an MLP instance can be precisely controlled. For the settings in Table 16, it takes up 1,154 weights on average, and the largest difference in size between Polygrid and MLP is less than 25 weights. The introduction of size variance by the control mechanism is due to the fact that, for some target sizes, the induced number of neurons in the hidden layer is fractional. As a consequence, many instances will be smaller and others larger.

The DT model occupies an intermediate position in the control spectrum. The strategy, which is based on restricting the depth of its tree, is less precise because the size of a decision tree depends on both its hyperparameters and the distributional properties of the data. In practice, the DT model takes 400 weights on average for the settings in Table

[15] Here, "model size" refers to the average number of weights taken up by instances of a model during its evaluation. Refer to Section 4.5.1 for a detailed description of the concept.

Figure 27 – Results from a Polygrid instance in a label ranking task

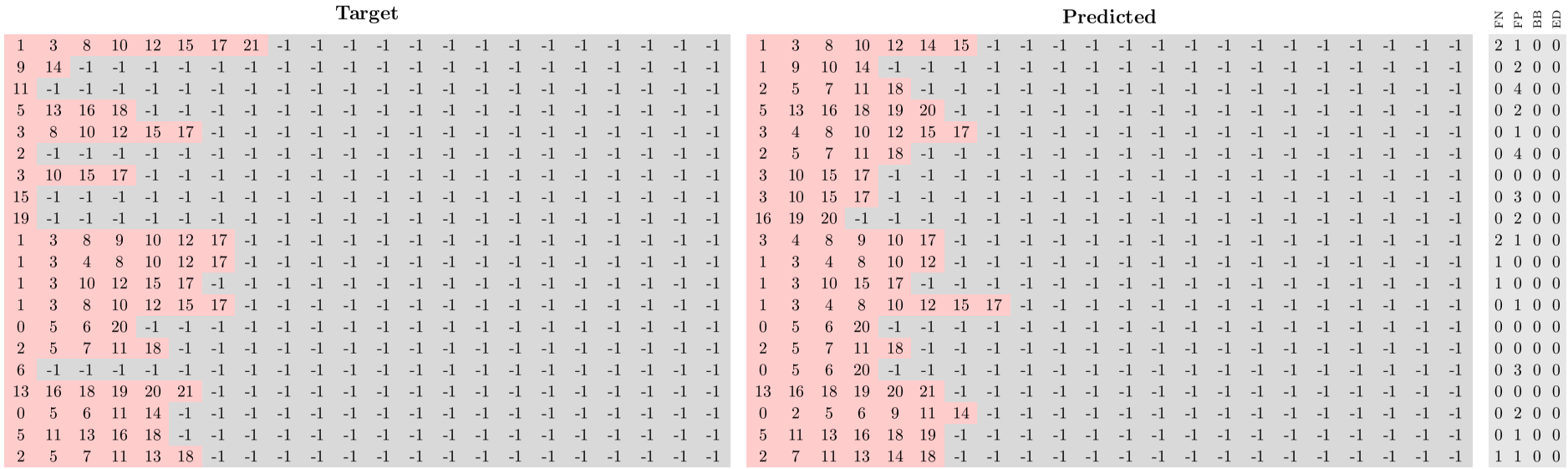


Target																						Predicted																						FN	FP	BB	ED
12	3	10	15	8	1	17	21	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	12	3	10	-1	-1	17	-1	8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	3	0	2	1
14	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	14	9	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1	0	0
11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	11	5	7	-1	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	3	1	0
18	13	16	5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	13	16	-1	18	-1	-1	-1	19	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	2	1
15	12	3	10	8	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	12	15	3	8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	2	0	0	1
2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	7	11	5	2	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	4	0	0
15	17	3	10	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	17	15	10	3	12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1	0	2
15	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	15	17	12	3	10	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	4	0	0
19	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	16	-1	-1	-1	-1	8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	2	1	0
9	8	10	3	1	17	12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	9	8	10	-1	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	3	0	1	0
12	3	10	8	1	4	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	4	1	12	8	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	2	0	0	4
17	10	3	15	12	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	17	10	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	3	0	0	0
3	17	10	8	15	12	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	4	3	17	15	1	10	8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	0	2
6	0	20	5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	6	0	20	5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0
7	18	2	11	5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	7	5	-1	18	11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	0	1	2
6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	6	0	5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	2	0	0
18	21	13	16	20	19	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	21	13	16	18	20	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	0	0	1
5	11	0	14	6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	5	14	11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	0	0	2
13	18	5	16	11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	13	16	11	5	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	2
2	7	11	18	13	5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	7	14	2	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	3	1	0	1

Source: The author.

Legend: The matrix in the left (Target) reproduces the content of 20 cases in the test partition during Polygrid's performance evaluation for config 2757 on the whoqol-lr-22 dataset. This is the config in which Polygrid achieves its top subseq accuracy. We selected a random split in which the observed accuracy is close to the one reported in the results for that config (subseq accuracy about 0.05). The matrix in the middle (Predicted) reproduces the output generated by the trained Polygrid instance for the corresponding cases. Ideally, both matrices should be identical. Finally, the rightmost matrix shows a breakdown of the types of errors committed. The first column (FN) indicates the number of false negatives (labels that should have been included in the prediction, but have not). The second column (FP) indicates the number of false positives (labels that should not have been included in the prediction, but have). The third column (BB) shows the count of bubbles. This corresponds to the number of times a filler sequence is improperly introduced in the prediction. A bubble represents a label with a higher score than that of a subsequent label, but which does not reach the threashold to be predicted as a positive case. For example, in the first row of the Predicted matrix, there are two sequences of fillers: one between the labels 10 and 17, and one between 17 and 8. The latter filler represent the label 21, whose score and reference threshold are (0.101, 0.127), while the label 8 has (0.097, 0.077) as score and threshold. Finally, the last column (ED) indicates the number of edits needed to transform the predicted sequence into the target sequence. This number does not account for operations needed to tackle false negatives and false positives.

Figure 28 – Results from a Polygrid instance in a mutilabel classification task



Target

1	3	8	10	12	15	17	21	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
9	14	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	13	16	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	8	10	12	15	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	10	15	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
15	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
19	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	3	8	9	10	12	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	3	4	8	10	12	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	3	10	12	15	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	3	8	10	12	15	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	5	6	20	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	5	7	11	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
13	16	18	19	20	21	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	5	6	11	14	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	11	13	16	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	5	7	11	13	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

Predicted

1	3	8	10	12	14	15	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	9	10	14	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	5	7	11	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	13	16	18	19	20	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	4	8	10	12	15	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	5	7	11	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	10	15	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	10	15	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
16	19	20	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	4	8	9	10	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	3	4	8	10	12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	3	10	15	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	3	4	8	10	12	15	17	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	5	6	20	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	5	7	11	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	5	6	20	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
13	16	18	19	20	21	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	2	5	6	9	11	14	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	11	13	16	18	19	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	7	11	13	14	18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

FN	FP	BB	ED
2	1	0	0
0	2	0	0
0	4	0	0
0	2	0	0
0	1	0	0
0	4	0	0
0	0	0	0
0	3	0	0
0	2	0	0
2	1	0	0
1	0	0	0
1	0	0	0
0	1	0	0
0	0	0	0
0	0	0	0
0	3	0	0
0	0	0	0
0	2	0	0
0	1	0	0
1	1	0	0

Source: The author.

Legend: The matrix in the left (Target) reproduces the content of 20 cases in the test partition during Polygrid's performance evaluation for config 2757 on the whoqol-ml-22 dataset. The cases are the same ones used in the preceding analysis for label ranking (see Figure 27). Note that the data was converted from label presence to label ranking format to make comparisons easier. The matrix in the middle (Predicted) reproduces the output generated by the trained Polygrid instance for the corresponding cases. The rightmost matrix shows a breakdown of the types of errors that were committed. The first column (FN) indicates the number of false negatives (labels that should have been included in the prediction, but have not). The second column (FP) indicates the number of false positives (labels that should not have been included in the prediction, but have). The remaining columns represent issues that do not occur in multilabel classification tasks.

16. This means the DT model faces a problem similar to the Linear model's: it cannot grow beyond a certain size. Except on the multiclass datasets, the control mechanism did not exert a limiting effect on the sizes of instances of the DT model. Therefore, the argument should be dismissed for the DT model on the multilabel and label ranking datasets. However, a similar analysis applied to the remaining tree-based models (RF, BRRF, and BRDT) leads to a different conclusion. In general, these models can grow larger than Polygrid instances in any given setting. This means that the control mechanism exerts a limiting effect on their instances, and this may amplify their size variance.

In summary, we conclude that the argument about bias can be dismissed for the Polygrid, Linear, Ridge and DT models, but cannot be dropped for the MLP, RF, BRRF, and BRDT. Fortunately, the fact that the size of an MLP instance can be precisely controlled allows us to attempt a more rigorous response to the argument, focused on the MLP model. The response consists of a statistical argument that goes like this: (a) assume that the bias in question exists and it is strong, (b) deduce how the bias would manifest in the data, and (c) use the data to assess the plausibility of the initial assumptions.

Accordingly, let's articulate the assumptions. Provisorily assume that (P1) the performance of the MLP model is superior to Polygrid's for a given setting[16], (P2) whenever the control mechanism is active, it negatively biases MLP's performance readings, and (P3) this bias is strong enough to be detected. The following claims are then granted:

- For a target size that induces an integer number of neurons in the hidden layer, the control mechanism is inactive. Thus, we have $var(\text{size}_{MLP}) = 0$ and the score perf_{MLP} is unbiased. Then, by premise P1, we have $\text{perf}_{MLP} - \text{perf}_{Ply} > 0$;
- For a target size that induces a fractional number of neurons, the control mechanism is active. Thus, we have $var(\text{size}_{MLP}) > 0$ and, by premises P2, the score perf_{MLP} is negatively biased. Then, by premise P3, we have $\text{perf}_{MLP} - \text{perf}_{Ply} \leq 0$;

Consequently, the prototype pattern shown in Figure 29a should be the predominant pattern found in the performance data, as it characterises how the bias would manifest.

To extract these patterns from the data, the first step is to find all configs in the evaluation's parameter space such that the size of a Polygrid$(n_s/d, n_a)$ instance induces an integer number of neurons in the hidden layer of an MLP model[17]. For each of these configs, execute the pipeline illustrated in Figure 29a. The leftmost matrix shows values for Polygrid's hyperparameters. The central cell holds the special tuple $(n_s/d, n_a)$, and

[16] By setting we mean a combination of dataset and a couple of Polygrid's hyperparameters, solver and cutoff. Recall that the size of a Polygrid instance depends on the dimensionality of the data, and the four hyperparameters $n_s/d, n_a$, solver and cutoff.

[17] To improve the clarity of the presentation, the other hyperparameters that affect Polygrid's size, namely solver and cutoff, are omitted in the remaining of this argument.

the surrounding cells hold variations of this tuple, created by increasing or decreasing each hyperparameter by one unit. The "target sizes" matrix shows the size of the Polygrid instances with the hyperparameters of the previous matrix. The "MLP" and "Polygrid" matrices show the performance scores these models obtain. It must be noted that the score in the centre of the "MLP" matrix does not carry any effects of the control mechanism, while the scores in the surrounding cells may carry variance introduced by it. Finally, the difference in performance (i.e., $\text{perf}_{MLP} - \text{perf}_{Ply}$) is shown in the "Difference" matrix, and the "Pattern" matrix is computed by extracting the sign of the difference scores.

The extracted patterns can be used to assess the plausibility of the premises. For example, the finding of a large proportion of patterns in which the central cell is marked with a plus sign would support premise P1. More importantly, a histogram of the Hamming distance between each extracted pattern and the prototype pattern should display a large concentration of mass near zero, lending support to premises P2 and P3 jointly.

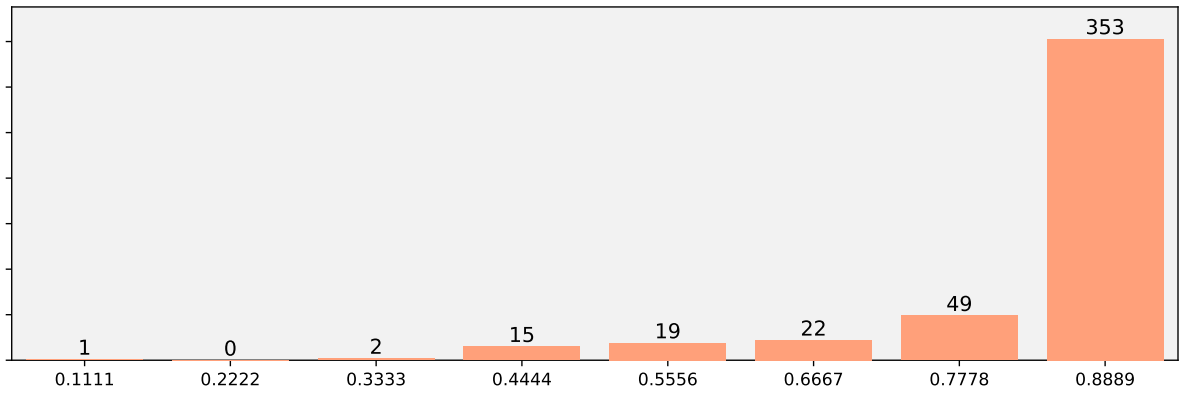
Applied to the performance data, the procedure just described identified 576 eligible configs: 216 configs involving multiclass datasets, 180 configs involving multilabel datasets, and another 180 for label ranking datasets. Among the extracted patterns, we did find a large proportion of patterns with the central cell marked with a plus sign, of about 80%. The patterns with a minus sign in the centre occur only for multiclass datasets and, since they do not satisfy premise P1, they cannot work as a contrast in this analysis and are discarded. The remaining 461 patterns with the central cell marked with a plus sign were used to plot the histogram shown in Figure 29b. Contrary to the expected distribution under premises P1 to P3, the mass is not concentrated near zero. In fact, the opposite distribution was observed: most extracted patterns strongly disagree with the expected pattern, as seen in Figure 29c. This result renders premises P2 and P3 jointly implausible, and we conclude that the control mechanism did not cause a negative bias in the evaluation of MLP's performance that is strong enough to be detected with our deductive device. Thus, the argument about bias should also be dismissed for the MLP model.

As a closing remark on the fairness of this evaluation, we want to say a few words about the argument that Algorithms 10 and 11 are arbitrarily coordinated. They implement an approach in which Polygrid is used an "anchor": the search space is defined in terms of Polygrid's hyperparameters, the best Polygrid configs are found, and then instances of the alternative models with comparable sizes are evaluated. In principle, two other paths are feasible: (a) to select another model as the anchor, or (b) to replace the use of an anchor with levels of a property shared by all models being evaluated. The first path was initially considered: the idea was to select the RF model because it appears in the top of rankings for both multilabel classification and label ranking tasks. The challenge was that that model produces a relatively large variance in the size of its instances, making comparisons less reliable, we feared. Regarding the second approach, we considered using the size of the

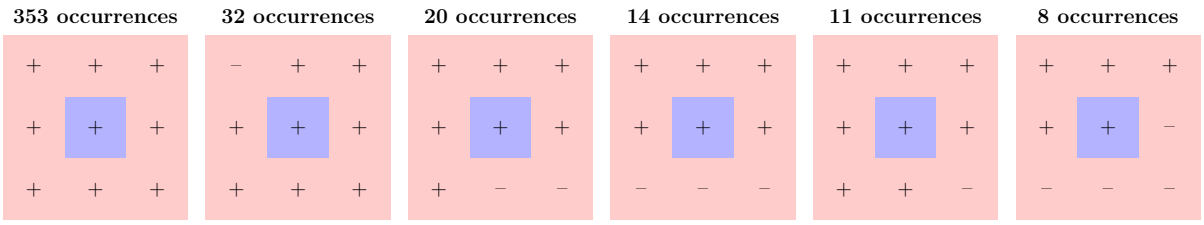
Figure 29 – Evidence against the evaluation's negative bias for the MLP model

(a) An example of the computation of the pattern matrix

(b) Histogram of distances between found patterns and the prototype pattern



(c) The six most frequent patterns extracted from the performance data



Source: The author.

Legend: The leftmost matrix in Figure 29a shows values of Polygrid's hyperparameters n_s/d and n_a. The hyperparameters in the blue cell induces an integer number of neurons h in the hidden layer of the corresponding MLP. The surrounding cells in red hold variations of this config, created by increasing or decreasing each hyperparameter by one unit. The hyperparameters in the red cells may induce fractional values of h. The "target sizes" matrix shows the size of a Polygrid instance with the corresponding hyperparameters. The MLP and Polygrid matrices hold the accuracy scores for these models across the target sizes. Assuredly, the score in the blue cell of the MLP matrix does not carry any effects of the size control mechanism. The Difference matrix is computed as (MLP - Polygrid), and the Pattern matrix extracts the sign of the scores of that matrix. The term "prototype pattern" refers to the example shown. Figure 29b shows a histogram of Hamming distances between extracted patterns and the prototype pattern for patterns satisfying premise P1. Finally, Figure 29c show the six most frequent patterns that were observed in the data.

models. The issue, however, was to justify the choice of the levels to be included in the search space. All in all, choosing Polygrid as an anchor lead to a better design, we believe.

5.3.5 Hints about how to choose hyperparameters for Polygrid

The Polygrid model has several hyperparameters that can be independently selected: the type of sectors and the number of sectors per domain (n_s/d), the number of annuli (n_a) and their type, the method to select the order in which the domains are mapped to vertices of the assessment polygon (vorder), the algorithm used to solve linear equations (solver), and the cutoff scheme. This variety gives the model flexibility to adapt to different contexts, but also makes model selection a challenging task.

During this project, we tested the Polygrid model on a wide range of configurations. Unfortunately, we found no generally good heuristics to select its hyperparameters. So, the recommended approach is the standard one: define a parameter space and search for good configurations in that space. However, some lessons we learned may be useful in defining the search space to maximise labelset accuracy or minimise label-level loss:

- The configs found in the first stage of this evaluation consistently described models with the maximum numbers of annuli and sectors in the search space specified in Section 5.1.4 — 3 sectors per domain and 7 or more annuli. This suggests that a larger range would reveal models with higher performance. The trade-off is the computational cost and a threat to the interpretability of the Polygrid instances.

- Regarding the order in which domains are mapped to vertices, two of the three options appeared around the same number of times: averages and measures. The remaining option, rho, appeared much less frequently. This suggests that trying to maximise the average area of the assessment polygons for a dataset — which is the heuristic pursued by the first two — was more effective than placing domains that correlate the highest in neighbouring vertices — the heuristic pursued by the latter.

- The two options of the cutoff scheme occurred with about the same frequency. Both should be kept in the search space. However, the choice of multiple cutoffs adds explanatory power to the scale diagrams. In fact, it might be difficult to justify a single cutoff being shared by all target scales from a conceptual point of view.

- Regarding the solver options, a closer look at the results suggests that both lstsquni and lstsqsym could be removed from the search space with minor impact. This is because of the five configs using the lstsquni solver, four find a competitive alternative in a config with lstsq solver. A similar case can be made for the lstsqsym solver: of the nine configs, seven find a competitive config based on the lstsq solver.

These recommendations reduce the search space from 3,456 configs to 1,728 configs, and increase the range of both the number of sectors per domain $(1 \ldots 4)$ and number of annuli $(2 \ldots 10)$. However, this smaller search space does not necessarily demand a smaller computational effort because new configs that specify larger models were introduced.

5.4 Summary and closing remarks

This chapter reported on the results of an offline evaluation of the Polygrid model. This was a challenging task because there are no benchmarks for our application domain: gerontological primary care. We worked with domain specialists to identify relevant tasks, and give shape to the collection of datasets we ended up using as benchmarks. Due to the importance of interpretability to the application, and its decrease as model instances grow in size, we adapted the standard evaluation methodology to enforce the comparison between models whose instances have the same size on average.

The results of the evaluation indicate that Polygrid dominated the alternative models on the multiclass datasets, achieved mixed results on the multilabel datasets, and was consistently dominated on the label ranking datasets. A practical implication of this outcome is that, if field applications are considered, the researcher or practitioner is well advised to start the project by tackling first the multilabel classification task.

Regarding our goal of identifying ideas to improve our model, we pointed out that structural similarities between Polygrid and both the DT model (a CART regression tree) and the MLP model (a perceptron with a single hidden layer) suggest that a significant increase in performance may be attained by making Polygrid's feature extraction an adaptive process. In its current design, Polygrid does not consider the information relating the assessment data to the target assignments, except when the option "tree" is selected for the hyperparameter that controls the partitioning of the unit disc into multiple annuli. Efforts in this direction would probably benefit from the development of a closed form of the feature extractor in Equation 4.2 because, in principle, this could enable the use backpropagation-like techniques to learn optimal placements of sectors and annuli.

Finally, some recommendations about how to select hyperparameters for the Polygrid model were made based on our experience during the making of this evaluation. We could not derive general heuristics that are useful for hyperparameter selection, but we described a suggestion to modify the search space to focus on more promising configs. It must be noted, however, that a key idea that should constrain the specification of this search space is related to interpretability. After all, how interpretable is a Polygrid instance in which the unit disc is partitioned into 100 cells? Unfortunately, we could not address this question in this work, but the next chapter provides some promising evidence that, for small instances, the Polygrid model is indeed highly interpretable to layman users.

6 A USER STUDY TO ASSESS THE POLYGRID DIAGRAM

The conceptual framework introduced in Section 4.5 defines interpretability as a property that emerges from the interaction among data, model, and users. According to that framework, the contribution that a model brings to interpretability comes from its transparency, scalability, and ability to preserve meaning. Applied to Polygrid, we concluded that the model has these properties, which makes it conducive to interpretability. In fact, the conceptual defence of the Polygrid's interpretability presented in that section focused on demonstrating that the instantiated forward computational graph of a recommendation is visually encoded in its respective explanation diagram, and that the model's architecture allows the elements of the computational graph inherit the meaning of the data.

In this chapter, we complement that conceptual defence with an empirical evaluation of the model's interpretability. We report on the results of a user study that combines methodological elements from the literature on the interpretability of machine learning models and visualisation research (Doshi-Velez; Kim, 2017; Saket; Endert; Demiralp, 2019): the participant is shown an explanation diagram in which the assessment chart displays measurements taken from a flower specimen. Of course, the explanation diagram is stripped of its tags. The participant is asked to classify that specimen into one of three species.

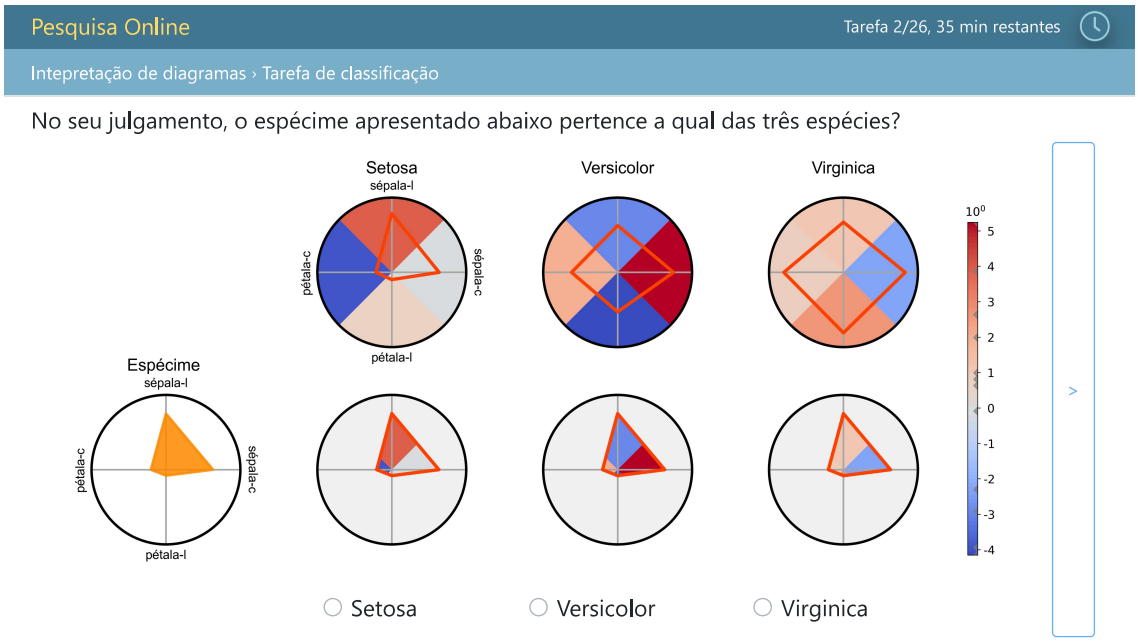
6.1 The design of the interpretability assessment user study

This study follows a within-subjects design in which each participant is exposed to two experimental conditions. In one condition, which we call "Polygrid condition", the participant is asked to perform a series of visual classification tasks using Polygrid diagrams, such as the one shown in Figure 30a. The diagrams were generated by training a Polygrid instance on the Iris dataset (Unwin; Kleinman, 2021). The task consists in classifying a specimen into one of the three species: Setosa, Versicolor, or Virginica. In the other condition, which we call "Barsgrid condition", the participant is asked to perform the same task, but using Barsgrid diagrams, which are identical to Polygrid's except for the radar charts being replaced by bar charts, as shown in Figure 30b. In addition to the participant's response, we collect the time spent to complete each visual classification task.

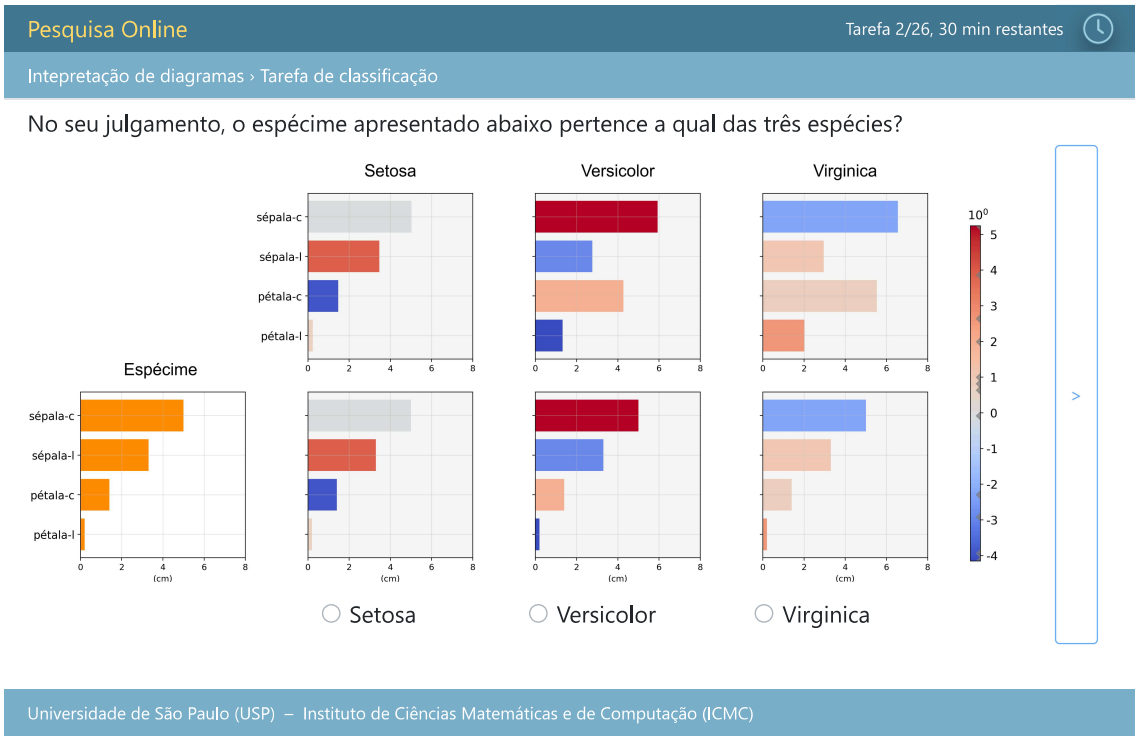
The collected data allow us to assess the interpretability of Polygrid diagrams by means of a proxy measure based on the definition adopted in Section 4.5: interpretability is the measure of success participants achieve when asked to perform forward simulation tasks in an experimental context. Moreover, the data also allow us to assess how Polygrid performs in comparison to the use of bar charts to support visual classification tasks. In the remainder of this section, we detail how the participant's journey is organised and how the collected data is used to assess the variables that are relevant to this study.

Figure 30 – Screenshots of the webpages used in the two experimental conditions

(a) Visual classification task in the Polygrid condition



(b) Visual classification task in the Barsgrid condition



Source: The author.

Note: Screenshots of webpages for the visual classification task under the Polygrid and Barsgrid conditions. Both diagrams display the case 49 of the Iris dataset. The textual elements are presented in Portuguese, and the main instruction reads “In your opinion, the specimen shown below belongs to which of the three species?” Before landing on this page, the participant watches a video with instructions on how to perform the task, including how to read the diagrams, record their judgement, and move to the next task.

6.1.1 Experimental conditions

How can we assess the interpretability of Polygrid diagrams by means of a user study in which the participant is asked to perform visual classification tasks? Is there a relationship between interpretability and visual classification tasks in our context? As said before, we adopted an operational definition of interpretability as the measure of success participants achieve when asked to perform forward simulation tasks in an experimental context. In this sense, a forward simulation task is a task in which "humans are presented with an explanation and an input, and must correctly simulate the model's output" (Doshi-Velez; Kim, 2017). In a Polygrid diagram, the assessment chart represents the input, and the explanation is provided by the assignment and matching charts. In this setting, success depends primarily on the participant's ability to assess the similarity of two visual shapes: the assessment polygon and the class prototypes. Thus, we take the participant's accuracy in reproducing the Polygrid's output as a proxy measure for its interpretability.

However, this method of evaluating interpretability creates the need to make some adaptations to the Polygrid diagram. The reader may have noticed that, in Figure 30a, the Polygrid diagram was stripped of an element that appears in the diagrams shown in earlier chapters: all tags were removed, including the tags informing the degree of similarity between the input and the labels. The tags were removed because they carry information about the output of the model, which is precisely what the participant must replicate. Furthermore, the interactivity of the Polygrid diagram was disabled, so that the participant could not inspect the value assigned to the elements of the diagram.

It is important to highlight that these adaptations preserve an essential feature of the Polygrid diagram: the use of polygons to represent multidimensional data. In principle, this representation drives a person to resort to their visual intuition (of shape sorting) instead of their analytical abilities (of comparing numerical scores). Actually, this choice of representation aims to nudge the participant's attention to focus on a holistic view of the latent variable represented by the area of the assessment polygon, rather than on the individual domain scores. The decision to include the Barsgrid condition in the study design was mainly driven by the need to investigate the presumed impact of this choice of visual representation on the participant's accuracy in solving the tasks.

The Barsgrid diagram was designed to be as similar to Polygrid's as possible. The substitution of radar charts for bar charts was based on results from the literature on the task-based efficiency of basic visualisations: empirical evidence suggests that bar charts are among the best basic visualisations to support visual classification tasks[1]. Moreover, the bar chart naturally drives the participant's attention to focus on domain scores individually. This creates the contrast needed to estimate the effect of using polygons to represent

[1] Based on the taxonomy of Amar, Eagan and Stasko (2005), our visual classification task can be described as a combination of filter and sort basic analytic tasks.

multidimensional data. Furthermore, the bar chart gracefully accommodates a scale shared by all domains. This feature helps us investigate the effect of the absence of numerical scales in Polygrid diagrams, which was pointed out as a deficiency during a pilot study.

As we just said, the Barsgrid diagram was designed to be as similar to Polygrid's as possible. The latter results from the instantiation, composition, and coordination of nine different graphic types[2], as described in Table 18. The Barsgrid diagram was carefully created to mirror Polygrid's in terms of the visual encoding used in each graphic type, with a few exceptions that result from the substitution of radar charts without numerical scales for bar charts with a numerical scale that is shared by all domains.

Finally, the Iris dataset was chosen because (a) all of its features are measured in the same scale (cm), which satisfies our need to introduce a single numerical scale in the Barsgrid diagram, and (b) it describes generally familiar objects (flowers), which eliminates the need to recruit specialists, as would be the case had we chosen a healthcare dataset. We trained a Polygrid$(1, 1)$ instance with "sector = cover" so that annular sectors can be bijectively mapped to bars in the corresponding charts. This simplifies the video instructions, which can guide the participant to focus primarily on the similarity of shapes[3].

6.1.2 Methodology

The journey of the participant during the study is described by a script. A script is made up of blocks, and a block is a sequence of tasks, as illustrated in Figure 31. The script starts by presenting information about the study being conducted, such as motivation, objectives, methods employed, and how the collected data will be managed and used. If the participant gives their consent, then he or she is redirected to a video that explains what is expected from the participant and gives instructions on how to perform the visual classification task corresponding to one of the experimental conditions (e.g., Polygrid condition). Once the instruction task is completed, the participant is asked to complete three visual classification tasks. These are warm-up tasks, and their objective is to verify that the participant understands how to perform the task explained in the video. After completing the warm-up tasks, the participant is shown a positive reinforcement message and is informed that he or she will be asked to complete eight new visual classification tasks that may be more difficult to solve than the ones just completed. Upon completion, the participant is submitted once again to the same sequence of tasks (starting from the video instruction), but now the tasks are specialised for the remaining experimental condition (e.g., Barsgrid condition). Finally, the script ends with a closing task, in which the participant is invited to provide contact and demographic information and is informed of the procedure to withdraw their data from the study, as required by local regulations[4].

[2] The term "graphic type" denotes a recurrent element of graphs, such as axes and labels.
[3] You can watch the instructional videos and try the visual classification tasks in this website.
[4] Item IV.3.d of the Resolução 466/2012 of the Conselho Nacional de Saúde - Brazil.

Table 18 – Breakdown of graphic types comprising Polygrid and Barsgrid diagrams

Graphic type	Role in the Polygrid diagrams	Role in the Barsgrid diagrams	Eq.
	Assessment Charts		
title	identifies the sample	identifies the sample	yes
domain label*	identifies the domain of an axis	identifies the domain of an axis	yes
outer bounds	confines the semiotically-charged elements of the chart, and encodes the upper limit of the domain scales	confines the semiotically-charged elements of the chart, and encodes the limits of the domain scales (both lower and upper limits)	yes
domain axis*	the interval within which the scores vary, **normalised** to the unit interval $(0, 1]$	the interval within which the scores vary, **without normalisation**	no
domain scale	absent	shows a scale shared by all domain axes	no
polygon (boundary)*	each chart contains a single polygon (the assessment polygon) whose vertices represent the normalised domain scores	each chart contains multiple polygons (the bars are rectangles), each representing a domain score of the sample being shown	no
polygon (inner filling)	single colour representing the type of the measurand (capacity or deficit)	single colour representing the type of the measurand (capacity or deficit)	yes
	Assignment Charts		
title	identifies the class label	identifies the class label	yes
domain label*	identifies the domain of an axis	identifies the domain of an axis	yes
outer bounds	same role as in the assessment chart	same role as in the assessment chart	yes
domain axis*	same role as in the assessment chart	same role as in the assessment chart	no
domain scale	absent	shows a scale shared by all domain axes	no
polygon (boundary)*	each chart contains a single polygon (the class prototype) whose vertices represent the average scores of samples of a class	each chart contains multiple polygons (bars), each representing a domain score averaged over samples of a class	no
polygon (inner filling)*	the colour that fills a region encodes its potential discriminative value in classifying samples of the respective class — actually, the colours in the background are not confined to the polygon.	the colour that fills a bar encodes the discriminative value that its respective domain has in classifying samples of the class	yes
polygon (outer filling)*		the colour corresponding to the null weight fills the space within the outer bounds and outside the bars (irrelevant for the task)	no
colour bar	maps colours to learned weights	maps colours and learned weights	yes
	Matching Charts		
title	absent	absent	yes
domain label	absent	absent	yes
outer bounds	same role as in the assessment chart	same role as in the assessment chart	yes
domain axis*	same role as in the assessment chart	same role as in the assessment chart	no
domain scale	absent	shows a scale shared by all domain axes	no
polygon (boundary)*	same role as in the assessment chart	same role as in the assessment chart	no
polygon (inner filling)*	the colour that fills a region encodes its contribution in determining the membership of the sample to the class	the colour that fills a bar encodes its contribution in determining the membership of the sample to the class	yes
polygon (outer filling)*	same role as in the assignment chart	same role as in the assignment chart	no
colour bar	same role as in the assignment chart	same role as in the assignment chart	yes

Source: The author.

Legend: Column "graphic type" lists a class of recurrent graphical elements (e.g., boundaries, axes, scales) that is instantiated to compose a concrete diagram. Column "Eq." indicates whether the graphic type conveys equivalent information in both diagrams. An asterisk after the name of a graphic type denotes multiplicity.

Before starting recruitment, 100 such scripts were generated and uploaded to the Marjory website, which can manage scripts in a format equivalent to that of Figure 31. The scripts were generated by training a Polygrid$(1, 1)$ instance on the Iris dataset split into 120 cases for training and 30 cases for testing. Only test cases were used to specify visual classification tasks. These cases were sorted according to their distance from the respective class prototype[5]. The three cases closest to each prototype were reserved for warm-up tasks (cases 49, 82, and 104). Of the remaining 27 cases, the first 13 closest to their respective prototype were assigned to the "left" pool, the 12 cases farther away from their prototype were allocated to the "right" pool, and the remaining two were reserved for mid and repeat tasks (cases 115 and 38). The cases in the left pool were randomly drawn to specify easy tasks, and the cases in the right pool were drawn to specify hard tasks.

Participants were invited by email containing a link to the website hosting the study. When a participant clicks on the link, the next free script in the study is allocated. To reduce order effects, scripts with an odd identifier (SID) specify the Polygrid condition as the first experimental condition to which the participant is exposed, and the Barsgrid condition as the last experimental condition. In contrast, scripts with an even identifier specify the opposite order. This is to say that some participants will start solving visual classification tasks using Polygrid diagrams, and others will start solving tasks using Barsgrid diagrams. In fact, the script with SID $= 2n$ with $(n = 1 \ldots 50)$ is a copy of the script with SID $= 2n - 1$ in which the order of the experimental conditions is reversed. Finally, the number of tasks (26 tasks) was chosen so that the time required to complete a script did not exceed 15 minutes, to avoid the possible effects of visual or cognitive fatigue.

6.1.3 Data analysis

The core variables of this study are the mean accuracy and the mean completion time. These variables are aggregated over distinct subsets of the collected data to help us assess the assumptions and relevant hypotheses of the study. These subsets correspond to data from the different blocks and conditions illustrated in Figure 31. For example, once the data are downloaded from the website hosting the study, the first step is to apply a set of exclusion criteria: scripts that were not completed are discarded, as well as completed scripts in which the participant failed the warm-up tasks. This ensures that the remaining data comes from participants who understood how to perform the tasks. Next, two other assumptions of the study design are evaluated. First, we seek evidence that the tasks are indeed arranged from easier to harder cases. This verification is important for two reasons: (a) the cases in the warm-up block were selected under the assumption that we have the ability to identify the easiest cases to classify, and (b) the cases drawn to assemble the go block are balanced with respect to difficulty, which supports the claim that the mean

[5] For example, since Case 49 is labelled as Setosa, its distance is computed from the prototype that appears in the assignment chart for the Setosa label (consult Figure 11 as reference).

Figure 31 – How the participant's journey is specified by a script

SID	Script Specifications																									
		First Experimental Condition													Last Experimental Condition											
	SSB	PIB	WUB			RSB	GOB								PIB	WUB			RSB	GOB						SCB
1	CPC	PC1	49	82	104	RST	103	75	115	38	41	131	115	38	BC2	49	82	104	RST	103	75	115	38	41	131	115	38	CPD
2	CPC	BC1	49	82	104	RST	103	75	115	38	41	131	115	38	PC2	49	82	104	RST	103	75	115	38	41	131	115	38	CPD
3	CPC	PC1	49	82	104	RST	95	84	115	38	51	98	115	38	BC2	49	82	104	RST	95	84	115	38	51	98	115	38	CPD
4	CPC	BC1	49	82	104	RST	95	84	115	38	51	98	115	38	PC2	49	82	104	RST	95	84	115	38	51	98	115	38	CPD

easiest | easy | mid | hard | repeat | easiest | easy | mid | hard | repeat

Source: The author.

Notes: A table describing four scripts. Each script specifies the journey of a participant during the study. A script is identified by a unique SID and is composed of blocks. Blocks are sequence of tasks. For example, the script with SID = 1 describes a journey that starts with the participant being presented with information about the study, and asked to confirm their consent (task CPC). Upon consenting, they watch a video with instructions for the first set of visual classification tasks (PC1 or BC1, depending on the experimental condition to which the participant is exposed first). Next, he or she is tasked with three warm-up visual classification tasks, which aim to assert understanding (tasks in the WUB block, numbers identify the dataset instance). The participant is shown a positive reinforcement message (task RST) and tasked with eight new visual classification tasks (in the GOB block). Once completed, the entire process is then repeated for the remaining experimental condition. In the end, the participant is invited to provide contact and demographic information (task CPD).

Legend: **Blocks**: script starting block (SSB), participant instruction block (PIB), warm-up block (WUB), ready-set block (RSB), go block (GOB), and script closing block (SCB). **Tasks**: collect participant's consent (CPC), collect participant's demographics (CPD), video instruction for Polygrid Condition First (PC1), video instruction for Barsgrid Condition First (BC1), video instruction for Polygrid Condition Last (PC2), video instruction for Barsgrid Condition Last (BC2), and ready-set (RST).

accuracy is meaningful. Second, we look for signs that participants may have experienced fatigue during the experiment. The effects of fatigue can confound with the core study variables, and additional caution is advised if its occurrence is detected.

Finally, the proxy measure for interpretability and the relevant hypotheses are assessed. The latter aim to clarify whether the Barsgrid diagram is more effective than the Polygrid diagram in supporting the tasks in our experiment. The effectiveness is approached from multiple angles, looking not only for evidence from accuracy and completion time but also from the consistency and consensus of the group of participants. The consistency measure evaluates whether an individual participant produces the same response when asked to classify the same case a second time. This measure is computed using the responses for the mid and repeat cases in the go block (cases 115 and 38). The consensus measure evaluates the degree of agreement among all participants with respect to their responses to the same cases. It is calculated based on the responses given to cases in the easy, mid, and hard tasks that appear in the go block. We formalise consistency and consensus as the Cohen's κ statistic for intra- and inter-agreement for categorical data (Stemler, 2004):

$$\kappa(Y_a, Y_b) = \frac{\mathbb{E}\left[\!\left[y_{aj} = y_{bj}\right]\!\right] - \mathbb{E}\left[\!\left[u_{1j} = u_{2j}\right]\!\right]}{1 - \mathbb{E}\left[\!\left[u_{1j} = u_{2j}\right]\!\right]} = \frac{p_o - p_e}{1 - p_e}, \tag{6.1}$$

where y_{aj} is the label assigned to case j by the annotator A, u_{ij} is a label randomly assigned to case j, and $[\![\cdot]\!]$ is the Iverson bracket. Thus, p_o is the observed agreement between the annotators and p_e is the expected agreement if the labels were assigned randomly.

6.2 Results

Individuals attending regular courses at the Universidade de São Paulo (USP), Universidade Federal de São Carlos (UFSCar), and Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA) were invited by email. During the period from March 21 to May 28, 2024, a total of $N = 56$ people participated in our study. This group is composed of 40 men, 9 women, plus 7 individuals who chose not to inform their sex; they were aged 17 to 55 years (24 on average); there were 37 undergraduates and 13 graduate students, of varied backgrounds, mainly STEM, applied social sciences, and gerontology.

Of the 56 scripts completed by the participants, ten were discarded because the participant failed (at least one of) the warm-up tasks. Of the ten discarded scripts, six are Polygrid first and four are Barsgrid first. Although a relationship between failure to pass the warm-up tasks and the experimental condition seems implausible ($\chi^2(1) = 0.68$, $p = 0.41$), the device used by the participant seems to provide a better explanation. In six of the ten scripts discarded, the participant used a mobile device to access the study website[6], against the recommendation given in the invitation. Considering that the proportion of participants who used a mobile device was about 14% (8 of the 56 participants), the relationship seems likely ($\chi^2(1) = 6.57$, $p < 0.015$). It should be added that the proportion of men and women in the study did not change after the exclusion criteria were applied.

Of the 46 scripts accepted, 24 are paired (i.e., pairs whose SID are $2n$ and $2n - 1$) and 22 are single. Of the latter, 9 are Polygrid first, and 13 are Barsgrid first. As will soon be shown, the presence of single scripts and the small imbalance between the study conditions do not threaten the design or the results of the study. Each case from the test partition appears in at least one script, and the number of occurrences by label is closely balanced (332 Setosa, 322 Versicolor, and 358 Virginica). The results of the data analysis are summarised in Table 20, in which the relevant hypotheses are evaluated at various levels of confidence. However, before we discuss whether the main hypotheses are supported by the data, the status of two design assumptions must be clarified.

First, let's approach the idea that the tasks are arranged in increasing order of difficulty within the scripts. The tests whose results are reported in row "A2a" of Table 19 are based on two confidence intervals: one around the mean accuracy in easy tasks (μ_{acc}^{easy}), and the other around the mean accuracy in hard tasks (μ_{acc}^{hard}). To reduce a possible confounding with fatigue, only data from tasks in the first experimental condition are considered. When the intervals do not overlap, the inequality encoded in the statement of the assumption is evaluated ($\mu_{acc}^{easy} > \mu_{acc}^{hard}$), otherwise the test is declared inconclusive. At the 80% confidence level ($\alpha = 0.20$), the statement is evaluated as true ($\mu_{acc}^{easy} = 0.957 \in [0.935, 0.989]$ and $\mu_{acc}^{hard} = 0.804 \in [0.750, 0.859]$, with 92 samples). However, this

[6] The user agent string collected during the consent confirmation contains the word "Mobile".

result masks a deeper contrast: constrained to the Barsgrid tasks, the statement also yields true, but it is inconclusive when constrained to the Polygrid tasks. This contrast suggests that the scale of difficulty employed by the mechanism that assembles the scripts produced clearly noticeable effects on participants solving the Barsgrid tasks, but less noticeable when solving the Polygrid tasks. In addition, it seems reasonable to assume that participants would take less time to complete easy tasks compared to hard ones. This assumption, though, is not backed by the completion time data from neither condition, as indicated in row "A2t" ($\mu_{time}^{easy} = 24.0 \in [21.3, 26.6]$ sec, and $\mu_{time}^{hard} = 17.6 \in [15.5, 19.7]$ sec with 88 and 87 samples respectively)[7]. Our hypothesis is that this systematic decrease is due to learning effects, as participants become more productive as they practice.

Second, let's seek for signs that the participants experienced fatigue during the experiment. When feeling fatigued, participants are less prone to meet cognitive demands that are perceived as low-stakes, as is usually the case in non-paid, voluntary participation in online studies. In principle, the larger the gap between demanded and volunteered efforts widens, the lower the accuracy and the less time spent on the tasks. In this case, a contrast between the participant's performance during the first and the last half of the study could unveil the effect of fatigue if it is strong enough. Moreover, we can constrain this analysis to the data from paired scripts to control for difficulty. This ensures that the same sequence of tasks appears as the first and last condition in a pair of scripts, as shown in scripts 1 and 2 in Figure 31. The remaining confounder, individual differences in ability between participants assigned to the Polygrid first or Barsgrid first scripts, seems to be implausible ($\chi^2(1) = 0.07, p = 0.79$). The results in row "A3a" suggest that there is no systematic decrease in accuracy between the first and last experimental conditions ($\mu_{acc}^{first} = 0.917 \in [0.889, 0.944]$ and $\mu_{acc}^{last} = 0.875 \in [0.840, 0.910]$, with 144 samples). However, constrained to Barsgrid tasks, a systematic decrease is observed at the 80% confidence level, which it is not accompanied by a decrease in completion time that would be expected in the event of fatigue, as reported in the row "A3t" for Barsgrid tasks. Our hypothesis is that this decrease is due to a momentary resistance to meet a cognitive demand that is perceived as larger than previously experienced when solving Polygrid tasks[8]. This point is illustrated in Figure 32. The top left chart (accuracy in Barsgrid tasks) shows a lower accuracy in easy tasks of the second condition compared to the first condition (Steps 19-20 vs. 6-7), and a comparable accuracy in the subsequent mid/repeat (Steps 8-9,12-13 vs. 21-22, 25-26) and hard tasks (Steps 10-11 vs. 23-24). All things considered, we assume that the design assumptions have been reasonably secured.

[7] We discarded 13 samples with a completion time larger than 90 seconds as outliers.

[8] This hypothesis is based on ideas from dual-process theories of decision-making (Evans; Stanovich, 2013). Participants rely on visual intuition to solve Polygrid tasks but must switch to more effortful analytical thinking for Barsgrid tasks. Because visual intuition is perceived as requiring less effort, participants who start with the Polygrid tasks are likely to experience an increase in cognitive demand when they move on to solve the Barsgrid tasks.

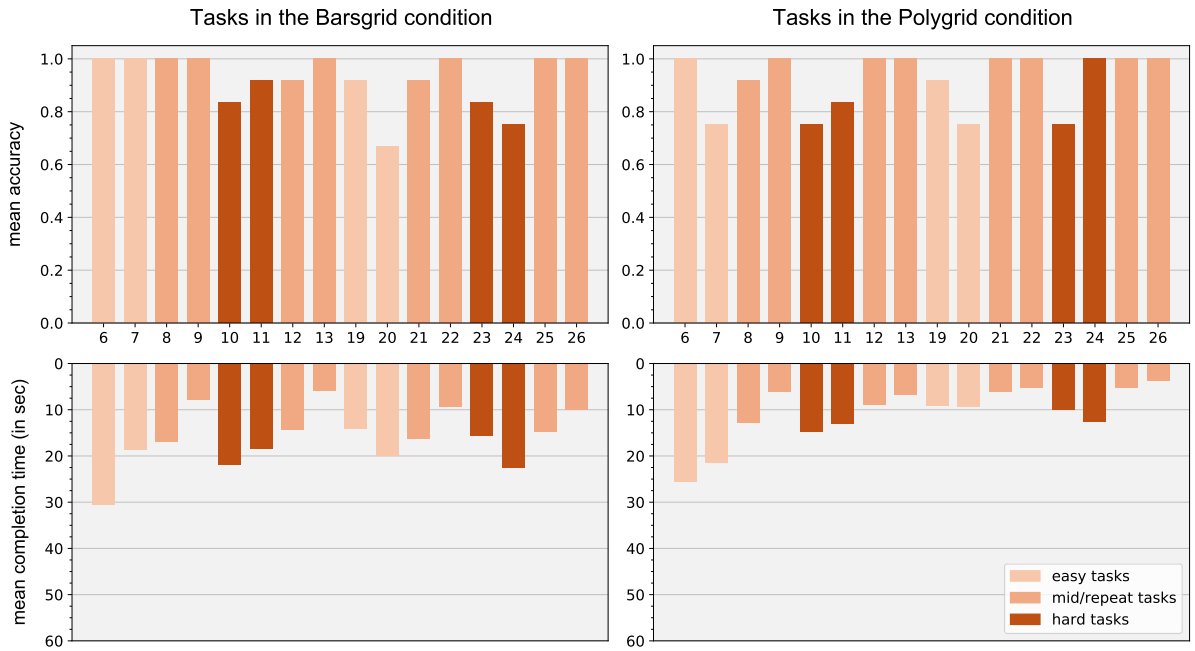
Table 19 – Design assumptions evaluated on data from the accepted scripts

ID	Statement of the Assumption	Confidence Level (α)				Estimate
		0.05	0.10	0.15	0.20	(at $\alpha = 0.20$)
A2a	Tasks are arranged in increasing order of difficulty (evidence from accuracy, only 1st condition)	True	True	True	True	92 samples
		mean accuracy in easy tasks				.935 .989
		in hard tasks				.750 .859
	— just Barsgrid tasks	True	True	True	True	50 samples
		mean accuracy in easy tasks				1.000
		in hard tasks				.720 .880
	— just Polygrid tasks	Inc	Inc	Inc	Inc	42 samples
		mean accuracy in easy tasks				.857 .952
		in hard tasks				.738 .881
A2t	Tasks are arranged in increasing order of difficulty (based on completion time, only 1st condition)	Inc	False	False	False	88 samples 87 samples
		mean time to complete easy tasks				21.3 26.6
		hard tasks				15.5 19.7
	— just Barsgrid tasks	Inc	Inc	Inc	Inc	47 samples
		mean time to complete easy tasks				22.5 30.0
		hard tasks				17.2 23.8
	— just Polygrid tasks	Inc	Inc	False	False	41; 40 smp
		mean time to complete easy tasks				17.8 24.7
		hard tasks				12.2 16.3
A3a	There is a decrease in accuracy (1st vs. 2nd condition, only paired scripts)	Inc	Inc	Inc	Inc	144 samples
		mean accuracy in 1st cond.				.889 .944
		in 2nd cond.				.840 .910
	— just Barsgrid tasks	Inc	True	True	True	72 samples
		mean accuracy in 1st cond.				.931 .986
		in 2nd cond.				.792 .903
	— just Polygrid tasks	Inc	Inc	Inc	Inc	72 samples
		mean accuracy in 1st cond.				.833 .931
		in 2nd cond.				.861 .944
A3t	There is a decrease in completion time (1st vs. 2nd condition, only paired scripts)	Inc	True	True	True	190 samples 192 samples
		mean completion time in 1st cond.				13.9 16.6
		2nd cond.				10.5 12.6
	— just Barsgrid tasks	Inc	Inc	Inc	Inc	95; 96 smp
		mean completion time in 1st cond.				14.7 18.9
		in 2nd cond.				13.6 17.1
	— just Polygrid tasks	True	True	True	True	95; 96 smp
		mean completion time in 1st cond.				11.9 15.5
		in 2nd cond.				6.8 8.6

Source: The author.

Note: In assumption A2a, the statement clarifies the scope of the assumption. To its right, the results of its evaluation at various levels of confidence are shown, followed by the number of observations in its scope. Two scales are plotted below the statement. They depict the confidence intervals (CIs) being contrasted: one around the mean accuracy for easy tasks, and the other around the mean accuracy for hard tasks. Description and details of each CI appear to the right of each scale. An opaque rectangle depicts a CI for $\alpha = 0.20$, and a transparent one depicts a CI for $\alpha = 0.05$. If the CIs overlap, they appear in blue, and red otherwise. Remaining rows are similarly organised. “Inc” stands for inconclusive, and “smp” for samples.

Figure 32 – Distribution of the core variables over the steps of the paired scripts



Source: The author.

Note: Four charts showing the core variables distributed over the steps of the paired scripts. Charts in the top row are based on participant's response data, and the ones in the bottom row are based on completion time; the ones at the left column show results from Barsgrid tasks, and the ones at right show results from Polygrid tasks. The scale for completion time is reversed to highlight its reciprocal relationship with accuracy. The steps are indicated in the abscissa, in chronological order, as described in Figure 31. Steps 6-13 correspond to tasks in the first experimental condition, and steps 19-26 to tasks in the last condition.

6.3 Discussion

Our proxy measure for the interpretability of Polygrid diagrams is the degree of success that participants achieve in reproducing the output of an instance using the diagram drawn for some case, as shown in the visual classification task in Figure 30a. For the sake of clarity, the instructional videos guided the participant to base their judgement primarily on the similarity between polygons and rely on the weights given to different regions as a tiebreaker. This is in tension with the Polygrid model, whose output is based on the weighted areas of the matching polygons. The results suggest that this difference was irrelevant: the overall mean accuracy of participants considering the Polygrid's output as the ground truth is statistically indistinguishable from the same estimate based on the original dataset labels as ground truth (respectively, $0.927 \in [0.910, 0.946]$ and $0.938 \in [0.921, 0.954]$, with 368 samples). In our view, this is evidence that Polygrid diagrams were easily interpreted by the study participants. We believe that this conclusion can be generalised to other groups: since the critical skill to perform the tasks in this study is (basic) graph comprehension, which is covered by the curricular guidelines for secondary education in Brazil (and elsewhere), the

general population would probably find the diagram interpretable. However, generalising to other contexts (e.g., other datasets, other Polygrid configs) will require further research, as we briefly discuss in Section 6.4.

We now move on to assess the relative effectiveness of the Polygrid and Barsgrid diagrams. With respect to accuracy, the results are inconclusive, as summarised in row "H1a" in Table 20 ($\mu_{acc}^{Bars} = 0.909 \in [0.888, 0.931]$ and $\mu_{acc}^{Poly} = 0.917 \in [0.895, 0.938]$, with 276 samples). In other words, the participant's accuracy when using the Polygrid diagram is comparable to that when using the Barsgrid diagram. This is evidence that the use of polygons to represent multidimensional data and the absence of a numerical scale in Polygrid diagrams is not associated with a decrease in the participant's accuracy in solving the tasks. However, it must be added that people made different mistakes when using one diagram or another. For instance, among the eight participants shown diagrams for case 103, everyone correctly classified the case with the Barsgrid diagram, but only four were able to do so using the Polygrid diagram. On the other hand, among the nine participants shown diagrams for case 52, everyone correctly classified the case with the Polygrid diagram, but only five were able to do so using the Barsgrid diagram.

Regarding the completion time, the collected data paint a different picture, as reported in row "H1t" in Table 20. On average, it took substantially less time to complete a task with a Polygrid diagram than with a Barsgrid diagram ($\mu_{time}^{Bars} = 18.8 \in [17.5, 20.0]$ sec and $\mu_{time}^{Poly} = 11.5 \in [10.6, 12.4]$ sec, with 267 and 272 samples, respectively). Overall, the results indicate that the participants completed the tasks faster and with at least the same level of accuracy when using the Polygrid diagram compared to the Barsgrid diagram. This is in agreement with comments from several participants who, after finishing the study, found the Polygrid diagram easier to use. In fact, nine out of the 15 people who provided feedback explicitly stated a preference for the diagram, citing its ease of use for the tasks. This finding is consistent with the idea that Polygrid diagrams drive the participants to rely on visual intuition, which is perceived as a less cognitively demanding strategy than the analytical reasoning required by Barsgrid diagrams.

In terms of consistency, equivalent results were observed for both diagrams, as reported in row "H2" in Table 20. When presented a case for the second time, most of the participants provided the same response as previously. Based on the results of a pilot study, cases 115 and 38 were found to be mildly discriminative in terms of the ability of the participants to solve the tasks. These cases were selected because our aim was to estimate consistency considering cases of average difficulty. However, case 38 turned out to be non-discriminative, as illustrated in Figure 33. Ideally, these cases would be positioned between the easy cases (the ones that all participants were expected to classify correctly) and the hard cases. These results suggest that a significant difference in consistency between diagrams is more likely to be found for tasks with higher levels of difficulty.

Table 20 – Relevant hypotheses evaluated on data from the accepted scripts

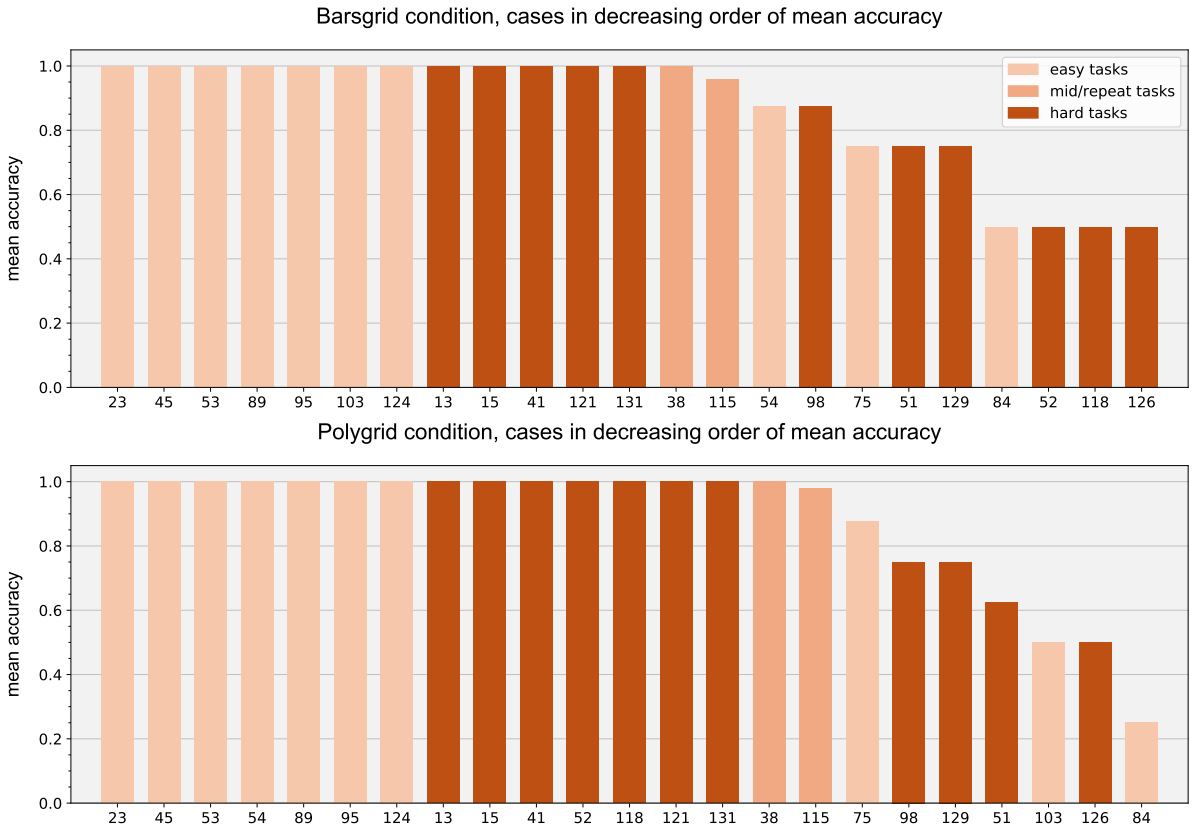
ID	Statement of the Hypothesis	Confidence Level (α)				Estimate
		0.05	0.10	0.15	0.20	(at $\alpha = 0.20$)
H1a	Barsgrid is more effective than Polygrid (evidence from accuracy, except repeat tasks)	Inc	Inc	Inc	Inc	276 samples
		mean accuracy in Barsgrid tasks				.888 .931
		in Polygrid tasks				.895 .938
	— just 1st condition	Inc	Inc	Inc	Inc	150;126 smp
		mean accuracy in Barsgrid tasks				.900 .953
		in Polygrid tasks				.865 .929
	— just 2nd condition	Inc	Inc	Inc	Inc	126;150 smp
		mean accuracy in Barsgrid tasks				.857 .929
		in Polygrid tasks				.907 .960
H1t	Barsgrid is more effective than Polygrid (evidence from completion time, all but repeat)	False	False	False	False	267 samples 272 samples
		mean comp. time, Barsgrid tasks				17.5 20.0
		Polygrid tasks				10.6 12.4
H2	Barsgrid is more effective than Polygrid (evidence from intra-agreement, just mid/repeat)	Inc	Inc	Inc	Inc	46 samples
		mean consistency, Barsgrid tasks				.942 1.000
		in Polygrid tasks				.971 1.000
H3	Barsgrid is more effective than Polygrid (evidence from inter-agreement, all but repeat)	False	False	False	False	1,035 smp
		mean consensus, Barsgrid tasks				.869 .887
		in Polygrid tasks				.898 .914

Source: The author.

Note: In assumption H1a, the statement clarifies the scope of the assumption. To its right, the results of its evaluation at various levels of confidence are shown, followed by the number of observations in its scope. Two scales are plotted below the statement. They depict the confidence intervals (CIs) being contrasted: one around the mean accuracy in Barsgrid tasks, and the other around the mean accuracy in Polygrid tasks. Description and details of each CI appear to the right of each scale. An opaque rectangle depicts a CI for $\alpha = 0.20$, and a transparent one depicts a CI for $\alpha = 0.05$. If the CIs overlap, they appear in blue, and red otherwise. Remaining rows are similarly organised. "Inc" stands for inconclusive.

Finally, with respect to consensus, the results suggest a systematic advantage of the Polygrid diagrams. More precisely, participants provided the same response to a given case more frequently when using Polygrid diagrams compared to Barsgrid diagrams ($\mu^{Bars}_{agree} = 0.878 \in [0.869, 0.887]$ and $\mu^{Poly}_{agree} = 0.906 \in [0.898, 0.914]$, with 1,035 samples). The large sample size is due to the fact that this measure is computed in an all-versus-all fashion: the responses given by a pair of participants to the set of cases that both evaluated are compared and a single agreement score is computed. Since all participants evaluate at least two common cases (the mid cases), there are $\binom{46}{2} = 1,035$ such pairs. Our hypothesis is that this finding is related to individual differences in reasoning strategies. In other words, if people vary less in their ability to match visual shapes than they do in handling numerical data, then the observed results would be a logical consequence.

Figure 33 – Average accuracy over the cases in the paired scripts



Source: The author.

Note: Two charts showing the mean accuracy over the cases of the paired scripts. The top chart shows mean accuracy in Barsgrid tasks, and the bottom one shows mean accuracy in Polygrid tasks. Cases indicated in the abscissa appear in decreasing order of mean accuracy. Multiple cases with the same mean accuracy are additionally ordered by type of task (easy, hard, and mid/repeat) and then by the number of the case. Case 115 is the first with a mean accuracy lower than one in both conditions. Ideally, all cases at its left should have been allocated to the left pool (i.e., the set of cases drawn to create easy tasks) and all cases to its right should have been allocated to the right pool (for hard tasks).

6.4 Concerns and limitations

The design of this study is based on the idea that interpretability depends on data, model, and user in intricate ways. Thus, if one needs to assess the contribution of a model to interpretability in an experimental setting, then the other variables should be carefully manipulated so that the study results, poor or good, could be solely ascribed to the model. According to the conceptual framework introduced in Section 4.5, this means that we should employ low dimensional data that is meaningful to the candidate participants.

Initially, we considered using the ampiab-ml-11 dataset in this study because it is a healthcare dataset with real assignments. This would allow us to recruit specialists in the care of older people to participate in the study. However, in our view, this dataset has

three issues that discourage its use in user studies to assess interpretability. The first issue concerns the huge imbalance of the dataset, as reported in Table 12. The majority class (aka most frequent label), which gathers individuals who were referred to specialised care after a home consultation, occurs 86 times, while there are seven labels that occur less than five times, and four labels that occur only once. Moreover, of the 15 labelsets observed in the assignments, eight occur a single time. In a user study in which participants are asked to perform decision-making tasks[9], a balanced dataset is highly preferable because we can compare the participants' responses against a uniform distribution to account for chance effects, as in Equation 6.1, among many other reasons.

The second issue concerns the level of abstraction of the interventions. For example, the majority class gathers patients who were referred to specialised care, but the specialty of care (e.g., physiotherapy, neurology, cardiology, etc.) is not available in the dataset. This makes it difficult to the researcher running the user study to hypothesise how participants use the assessment data to select interventions, and also hinders the ability of learning models to combine variance in the assessment data with the assignment data to select good predictive hypotheses. Relatedly, the third and last issue concerns apparent inconsistencies in the application of referral policies. For example, one can find individuals with high deficit in the oral health domain that have not been referred to the attention of the oral health team, and also find referrals to that team of patients with low deficit in oral health.

In view of these difficulties, we selected a different dataset. Our choice for the Iris dataset addresses the requirements for a low dimensional dataset ($d = 4$) with attributes that are meaningful to the general public, which simplifies recruitment. Moreover, the choice also addresses the need to set up a contrast to assess the effect on accuracy of endowing diagrams with scales (no scales vs. shared numeric scales), as mentioned earlier.

The second limitation concerns our choice of "model size". Earlier, we argued that some of the results were probably generalisable because they mainly depend on skills that people in general can learn (e.g., graph comprehension). However, we reserve judgement on generalising the reported results to contexts in which the diagrams have a more complex structure (e.g., multiple annuli, multiple sectors per domain, four or more labels). In other words, it is not clear whether the interpretability or the relative advantages of the Polygrid diagram, which stem from reframing a decision-making task into a kind of shape-sorting task, would be preserved in other contexts.

Finally, we also disclaim that the inclusion of the Barsgrid condition in this study did not produce results that are immediately applicable to bar charts. In Section 6.1.1, we state that the bar chart is one of the most effective visualisations for visual classification tasks, which underpins the design of the Barsgrid diagram. However, note that the task in our experimental setting differs markedly from that of Saket, Endert and Demiralp (2019).

[9] In the sense discussed in Section 3.3, following the typology of Brumar *et al.* (2025).

6.5 Summary and closing remarks

In agreement with the conceptual analysis developed in Section 4.5, the results of the user study reported in this section indicate that the Polygrid diagram is conducive to interpretability. In addition, the results show that the absence of graduated scales in the domain axes did not hinder the participants' performance in the decision-making task, and suggest that the use of polygons to present multidimensional data apparently leads to higher degree of consensus among participants. In fact, the removal of graduated scales and the use of polygons seem to have reduced the mean completion time compared to the alternative condition in which graduated scales and isolated scores are presented to the participant. It must be noted that the measures taken in the experimental setup to avoid the effects of fatigue and order effects seem to have been effective.

The decision to remove the graduated scales was motivated by our reading of the dual process theory applied to graph comprehension (Evans; Stanovich, 2013): the intention was to nudge the participants to rely on their visual skills and inhibit the use of analytical reasoning. According to that theory, the performance of visual tasks is less effortful than the deliberative thinking needed to process numeric information. Another expected effect, which was not approached in this study, is that relying on visual intuition may lead to improved trust in the system, but this remains to be shown in our setting.

We point out many limitations of this study, which stem from the fact that the study design explores the simplest decision-making task, which is equivalent to a multiclass problem, with the simplest configuration of the Polygrid model. Besides, the interactive resources of the Polygrid diagram, which allows the user to inspect the values of the underlying instantiated forward computational graph, were disabled. This was an initial effort to demonstrate the interpretability of the Polygrid model, and more research is needed to assess whether the reported results generalise to settings closer to the target application, namely a deployment as an assistive tool in gerontological primary care.

As a final thought, we would like to mention that the Polygrid diagram may be also useful in other settings in which one needs to illustrate data from the multiplication of small matrices. For example, it can be argued that Figure 18 shows the product between two matrices, $S_{3\times k}$ (the assessment charts) and $W_{4\times k}$ (the assignment charts), where the inner dimension k is "curled" into the unit disc. Similarly, the same visualisation can also be framed as a tensor $\boldsymbol{T}_{3\times 4\times k}$, whose tubes are depicted in the curled dimension.

7 CONCLUSION

This project started with the challenge of developing a recommender system for activities related to active ageing. From the beginning, two ideas rooted in the literature on gerontology were salient: the use of psychometric instruments to standardise comprehensive geriatric assessments (CGA) and the recurring and independent calls of many authors to represent the assessment results as a radar chart. The need for an expert-in-loop was clear, but not many works in recommender systems reported the use of data from standardised assessments, let alone the provision of explanations based on these data. In hindsight, two key insights set the project in motion (and changed its focus to primary health care).

The first insight involved a conceptual shift: reframing the radar chart not merely as a visualisation (or idiom, in the jargon of the visualisation research community), but as an image that can be computationally analysed. The Polygrid's learning pipeline performs two critical steps — partitioning the unit disc and decomposing the assessment polygon — that functionally resemble a (primitive) computer vision pipeline where image segmentation and feature extraction transform raw images into feature vectors. This insight enabled us to project low-dimensional assessment data into higher-dimensional spaces and leverage techniques from collaborative filtering and multilabel ranking in designing the pipeline.

The second insight revealed a fundamental connection: the congeneric model — the workhorse underlying psychometric instruments in the factor-analytic tradition — constrains specific geometric features of the polygon that appears in a radar chart. This constraint allows us to assign meaningful interpretation to the polygon's area, which is a cornerstone of our argument for the interpretability of the Polygrid model. Parallel to this, the requirements for transparency and interpretability imposed the remaining constraints needed to align the model's architecture and the design of the explanation diagram.

Therefore, the Polygrid model, introduced in Chapter 4, leverages the structure of psychometric data to generate recommendations and provide explanations. It performs multilabel ranking tasks and, in this work, we advance its application to assist gerontologists in the creation of personalised care plans in primary care. The explanations are provided in a diagrammatic format that encodes the forward computational graph of the model, making these explanations faithful and the model transparent. When the model fits the data well, the explanation converts a decision-making task into a shape-matching task.

Adding to the theoretical basis for Polygrid's learnability presented in Chapter 4, empirical evidence is presented in Chapter 5, which reports on the results of an offline evaluation of the Polygrid model. We worked with domain specialists to give shape to the collection of datasets we ended up using as benchmarks. We adapted the standard

evaluation methodology to enforce the comparison between models whose instances have the same size on average, due to the importance of interpretability to our application and its decrease as model instances grow in size. The results show that Polygrid dominated the competing models on the multiclass datasets, achieved mixed results on the multilabel datasets, and was consistently dominated on the label ranking datasets. An implication of this outcome is that, if field applications are considered, the researcher is advised to start the project by tackling first the multilabel classification task.

The conceptual defence for Polygrid's interpretability laid out in Chapter 4 is complemented with the empirical evidence in Chapter 6, which reports the results of a within-subjects user study that combines methodological elements from the literature on the interpretability of machine learning models and tasks-based effectiveness in visualisation research: the participant is shown an explanation diagram in which the assessment chart displays measurements taken from a flower specimen, and is asked to classify that specimen. The results show that the participants achieved a mean accuracy of about 0.9, suggesting that the Polygrid diagram is conducive to interpretability. The study also suggests that the use of polygons to display multidimensional data and the absence of graduate scales in the Polygrid diagram did not hinder the participants' performance; rather, it reduced the mean completion time and increased the consensus among participants.

Future work

The investigations we conducted in this work have many limitations that may become approachable in the near future, as more resources in health care become public, or be approached, as more work is done. In the following, we highlight the key limitations and point out the direction that represents the natural evolution of the proposed model.

From Chapter 5:

- Access to a sample of the dataset being collected by the ICOPE Brasil initiative (Ferriolli *et al.*, 2024). We originally planned to include a sample of this dataset in our analysis, but the project schedule for data collection became prohibitive. The major threat to the validity of our results is related to the use of synthetic assignments. We use datasets with real-world health assessments, and the assignments were created in a principled way, but there is no safe substitute for real measurements, of course.

- Independently of the new dataset, a promising candidate for improvement is the method that converts the assignment matrix Y for label ranking into the membership matrix U. The analysis of the gap in Polygrid's performance between classification and ranking tasks identifies differences between Algorithms 5 and 8 as potential causes, and highlights the role played by the logranks function (Eq. 4.1) in this gap.

From Chapter 6:

- Assuming access to a sample of the dataset mentioned earlier, we should consider a new user study with professionals who have been involved in that project (as they have been trained to administer the instrument for intrinsic capacity). It seems to be a good idea to drop the Barsgrid condition, and focus on a size contrast (smaller vs. larger model instances). In this new study, another task should be included in the end of the participants' journey to ask her to fill out questionnaires for tolerance for ambiguity and reaction to uncertainty. These new variables may be useful for explaining variance in the observed decision-making behaviour (Shashar *et al.*, 2023).
- Further investigation can be conducted using the available resources. For example, new studies can focus on the effects of increasing instance size or the number of labels in the experimental conditions. Besides, the forward simulation tasks could be replaced with counterfactual simulation tasks, as described in Doshi-Velez and Kim (2017): "humans are presented with an explanation, an input, and an output, and are asked what must be changed to change the method's prediction to a desired output." The results may reveal the ways people parse and interpret the diagram.

From Chapter 4:

- A closed form for the feature extractor in Equation 4.2. This would allow us to investigate whether a universal approximation theorem can be proved for Polygrid. This would probably require a closed form that is amenable to functional analysis, so the tools developed by Cybenko (1989) or Steinwart (2002) could be reused.
- An independent improvement to consider is the adaptive implementation of this feature extractor. The challenge is to learn to partition the input space (i.e., to design an alternative to Algorithm 3) that combines both input and target data with no sacrifice to the Polygrid's interpretability and transparency.

Thus, although we point out several opportunities for further work that can be addressed with existing resources, real progress in the direction of deploying the proposed model as an assistive technology is fundamentally in need of quality health care data.

REFERENCES

ALIBERTI, M. J. *et al.* Validating intrinsic capacity to measure healthy aging in an upper middle-income country: Findings from the ELSI-Brazil. **The Lancet Regional Health - Americas**, v. 12, p. 100284, 2022. ISSN 2667-193X. Available at: https://doi.org/10.1016/j.lana.2022.100284.

ALLALOUF, M. *et al.* Music recommendation system for old people with dementia and other age-related conditions. *In*: INSTICC. **Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies**. Setúbal, Portugal: SciTePress, 2020. Volume 5: HEALTHINF, p. 429–437. ISBN 978-989-758-398-8. Available at: https://doi.org/10.5220/0008959304290437.

AMAR, R.; EAGAN, J.; STASKO, J. Low-level components of analytic activity in information visualization. *In*: **IEEE Symposium on Information Visualization (INFOVIS)**. Minneapolis, USA: IEEE, 2005. p. 111–117.

ANDRADE, S. C. V. d. **Análise psicométrica da Avaliação Multidimensional da Pessoa Idosa na Atenção Básica (AMPI/AB)**. 2019. Dissertação (Mestrado) — Escola de Artes, Ciências e Humanidades, University of São Paulo, São Paulo, Brazil, 2019. Available at: https://doi.org/10.11606/D.100.2019.tde-25102019-171740.

ANDRADE, S. C. V. d. *et al.* Health profile of older adults assisted by the elderly caregiver program of health care network of the City of São Paulo. **Einstein (São Paulo)**, Instituto Israelita de Ensino e Pesquisa Albert Einstein, v. 18, p. eAO5263, 2020. ISSN 1679-4508. Available at: https://doi.org/10.31744/einstein_journal/2020AO5263.

ANGELINI, L. *et al.* The NESTORE e-Coach: Designing a multi-domain pathway to well-being in older age. **Technologies**, v. 10, n. 2, 2022. ISSN 2227-7080. Available at: https://doi.org/10.3390/technologies10020050.

ARANHA, F.; ZAMBALDI, F. **Análise fatorial em administração**. São Paulo, Brazil: Cengage Learning, 2008. ISBN 978-85-221-0629-5.

AZMI, A. K.; ABDULLAH, N.; EMRAN, N. A. A recommender system model for improving elderly well-being: A systematic literature review. **International Journal of Advances in Soft Computing and its Applications**, International Center for Scientific Research and Studies, Malaysia, v. 11, n. 2, p. 87–108, 2019. ISSN 2074-8523.

AZMI, A. K.; ABDULLAH, N. B.; EMRAN, N. A. A collaborative filtering recommender system model for recommending intervention to improve elderly well-being. **International Journal of Advanced Computer Science and Applications**, v. 10, n. 6, p. 131 – 138, 2019. Available at: https://doi.org/10.14569/ijacsa.2019.0100619.

BENAVOLI, A.; CORANI, G.; MANGILI, F. Should we really use post-hoc tests based on mean-ranks? **Journal of Machine Learning Research**, v. 17, n. 5, p. 1–10, 2016. Available at: http://jmlr.org/papers/v17/benavoli16a.html.

BENTLAGE, E. *et al.* Cocreating a mobile health app providing physical activity recommendations for older people living with parkinson disease or dementia:

User-centered pilot study. **JMIR Formative Research**, v. 9, 2025. Available at: https://doi.org/10.2196/51831.

BERMINGHAM, A. *et al.* Automatically recommending multimedia content for use in group reminiscence therapy. *In*: **Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare, Co-located with ACM Multimedia 2013**. New York, USA: ACM, 2013. (MIIRH), p. 49–58. Available at: https://doi.org/10.1145/2505323.2505333.

BESIK, S. I.; ALPASLAN, F. N. RHCS - A clinical recommendation system for geriatric patients. *In*: GADEPALLY, V. *et al.* (ed.). **Heterogeneous Data Management, Polystores, and Analytics for Healthcare**. Cham, Switzerland: Springer, 2019, (Lecture Notes in Computer Science, v. 11470). p. 115–132. ISBN 978-3-030-14177-6. Available at: https://doi.org/10.1007/978-3-030-14177-6_10.

BOGATINOVSKI, J. *et al.* Comprehensive comparative study of multi-label classification methods. **Expert Systems with Applications**, v. 203, p. 117215, 2022. ISSN 0957-4174. Available at: https://doi.org/10.1016/j.eswa.2022.117215.

BOLLEN, K. A. **Structural Equations with Latent Variables**. USA: John Wiley & Sons, 1989. 1-514 p. ISBN 0-471-01171-1.

BOLLEN, K. A.; PEARL, J. Eight myths about causality and structural equation models. *In*: MORGAN, S. L. (ed.). **Handbook of Causal Analysis for Social Research**. Dordrecht: Springer Netherlands, 2013. p. 301–328. ISBN 978-94-007-6094-3. Available at: https://doi.org/10.1007/978-94-007-6094-3_15.

BORSBOOM, D. **Measuring the mind: Conceptual issues in contemporary psychometrics**. New York, US: Cambridge University Press, 2005. ISBN 978-0-521-84463-5. Available at: http://doi.org/10.1017/CBO9780511490026.

BORSBOOM, D.; MELLENBERGH, G. J.; HEERDEN, J. V. The theoretical status of latent variables. **Psychological Review**, American Psychological Association, v. 110, n. 2, p. 203–219, 2003. Available at: https://psycnet.apa.org/doi/10.1037/0033-295X.110.2.203.

BORSBOOM, D.; MOLENAAR, D. Psychometrics. *In*: WRIGHT, J. D. (ed.). **International Encyclopedia of the Social and Behavioral Sciences (Second Edition)**. Second edition. Oxford: Elsevier, 2015. p. 418–422. ISBN 978-0-08-097087-5. Available at: https://doi.org/10.1016/B978-0-08-097086-8.43079-5.

BRADSHAW, S.; HOWARD, P. N. Challenging truth and trust: a global inventory of organized social media manipulation. **The Computational Propaganda Project**, 2018. Available at: https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2018/07/ct2018.pdf. Access at: 6/30/2024.

BRINKER, K.; FÜRNKRANZ, J.; HÜLLERMEIER, E. A unified model for multilabel classification and ranking. *In*: **Proceedings of the 2006 Conference on ECAI 2006: 17th European Conference on Artificial Intelligence August 29 – September 1, 2006, Riva Del Garda, Italy**. NLD: IOS Press, 2006. p. 489–493. ISBN 1586036424.

BRINKER, K.; HÜLLERMEIER, E. Case-based multilabel ranking. *In*: **Proceedings of the 20th International Joint Conference on Artifical Intelligence**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007. (IJCAI'07), p. 702–707.

BROWN, A. S. *et al.* National Institutes of Health Consensus Development Conference Statement: Geriatric assessment methods for clinical decision-making. **Journal of the American Geriatrics Society**, Wiley-Blackwell, UK, v. 36, n. 4, p. 342–347, 1988. Available at: https://doi.org/10.1111/j.1532-5415.1988.tb02362.x.

BRUMAR, C. D. *et al.* A typology of decision-making tasks for visualization. **IEEE Transactions on Visualization and Computer Graphics**, v. 31, n. 10, p. 8536–8551, 2025. Available at: https://doi.org/10.1109/TVCG.2025.3572842.

CAI, Y. *et al.* Health recommender systems development, usage, and evaluation from 2010 to 2022: A scoping review. **International Journal of Environmental Research and Public Health**, v. 19, n. 22, 2022. Available at: https://doi.org/10.3390/ijerph192215115.

CALDERÓN-BLAS, J. A. *et al.* Medical recommender systems: A systematic literature review. *In*: **2023 Mexican International Conference on Computer Science (ENC)**. IEEE, 2023. p. 1–8. Available at: https://doi.org/10.1109/ENC60556.2023.10508695.

CARRINGTON, A. **Kernel methods and measures for classification with transparency, interpretability and accuracy in health care**. 2018. Tese (Doutorado) — University of Waterloo, Ontario, Canada, 2018. Available at: http://hdl.handle.net/10012/13735.

CARVALHO, I. A. *et al.* **Operationalising the concept of intrinsic capacity in clinical settings**. Geneva, 2017. Available at: https://www.who.int/ageing/health-systems/clinical-consortium/CCHA2017-backgroundpaper-1.pdf. Access at: 1/8/2021.

CAVANAUGH, D. *et al.* Prospective evaluation of comprehensive geriatric assessments in multidisciplinary bladder cancer care and implications for personalized vulnerability phenotyping. **Urologic Oncology: Seminars and Original Investigations**, v. 43, n. 8, p. 468.e7–468.e18, 2025. ISSN 1078-1439. Available at: https://doi.org/10.1016/j.urolonc.2025.03.025.

CHARTE, F. *et al.* Tips, guidelines and tools for managing multi-label datasets: The mldr.datasets R package and the Cometa data repository. **Neurocomputing**, 2018. ISSN 0925-2312. Available at: https://doi.org/10.1016/j.neucom.2018.02.011.

CHENG, W.; HÜHN, J.; HÜLLERMEIER, E. Decision tree and instance-based learning for label ranking. *In*: BOTTOU, L.; LITTMAN, M. (ed.). **Proceedings of the 26th International Conference on Machine Learning (ICML-09)**. Montreal, Canada: Omnipress, 2009. p. 161–168.

CHEUNG, K. L. *et al.* How recommender systems could support and enhance computer-tailored digital health programs: A scoping review. **Digital Health**, v. 5, 2019. Available at: https://doi.org/10.1177/2055207618824727.

COLLEN, M. F. *et al.* Automated multiphasic screening and diagnosis. **American Journal of Public Health and the Nations Health**, v. 54, n. 5, p. 741–750, 1964. Available at: https://doi.org/10.2105/AJPH.54.5.741.

COLLINS, G. S. *et al.* TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. **BMJ**, BMJ

Publishing Group Ltd, v. 385, 2024. Available at: https://www.bmj.com/content/385/bmj-2023-078378.

CROON, R. de *et al.* Health recommender systems: Systematic review. **Journal of Medical Internet Research**, v. 23, n. 6, 2021. Available at: https://doi.org/10.2196/18035.

CYBENKO, G. V. Approximation by superpositions of a sigmoidal function. **Mathematics of Control, Signals and Systems**, v. 2, n. 4, p. 303–314, 1989. Available at: https://doi.org/10.1007/BF02551274.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, v. 7, n. 1, p. 1–30, 2006. Available at: http://jmlr.org/papers/v7/demsar06a.html.

DHIVAKAR, S. Machine learning based cancer disease prediction. *In*: **International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)**. IEEE, 2022. p. 1–5. Available at: https://doi.org/10.1109/ICSES55317.2022.9914247.

DIEGOLI, H. *et al.* **Atlas de Variação em Saúde Brasil**. São Paulo, Brasil: Academia VBHC, 2022. ISBN 978-65-997447-0-9. Available at: https://www.academiavbhc.org/atlas. Access at: 9/19/2025.

DIGREGORIO, M. R. When everything goes wrong make a diagram. *In*: **Contributions from the Communicating Complexity Conference, 2CO Tenerife 2017**. Servicio de Publicaciones de la Universidad de La Laguna, 2020. Available at: https://doi.org/10.25145/b.2COcommunicating.2020.006.

DIMARA, E.; STASKO, J. A critical reflection on visualization research: Where do decision making tasks hide? **IEEE Transactions on Visualization and Computer Graphics**, v. 28, n. 1, p. 1128–1138, 2022. Available at: https://doi.org/10.1109/TVCG.2021.3114813.

DOSHI-VELEZ, F.; KIM, B. **Towards A Rigorous Science of Interpretable Machine Learning**. 2017. Available at: https://arxiv.org/abs/1702.08608.

DUTILH-NOVAES, C. Diagrams as joint epistemic actions: A dialogical account of diagrams in mathematical proofs. **The Journal of Mathematical Behavior**, v. 80, p. 101271, 2025. ISSN 0732-3123. Available at: https://doi.org/10.1016/j.jmathb.2025.101271.

ELSWEILER, D. *et al.* Second Workshop on Health Recommender Systems (HealthRecSys). *In*: SIGCHI. **Proceedings of the 11th ACM Conference on Recommender Systems**. New York, USA: ACM, 2017. (RecSys), p. 374–375. Available at: https://doi.org/10.1145/3109859.3109955.

ELSWEILER, D. *et al.* Engendering Health with Recommender Systems. *In*: SIGCHI. **Proceedings of the 10th ACM Conference on Recommender Systems**. New York, USA: ACM, 2016. (RecSys), p. 409–410. Available at: https://doi.org/10.1145/2959100.2959203.

ELSWEILER, D. *et al.* Third International Workshop on Health Recommender Systems (HealthRecSys). *In*: SIGCHI. **Proceedings of the 12th ACM Conference on Recommender Systems**. New York, USA: ACM, 2018. (RecSys), p. 517–518. Available at: https://doi.org/10.1145/3240323.3240336.

ELSWEILER, D. *et al.* Fourth International Workshop on Health Recommender Systems (HealthRecSys). *In*: SIGCHI. **Proceedings of the 13th ACM Conference on Recommender Systems**. New York, USA: ACM, 2019. (RecSys), p. 554–555. Available at: https://doi.org/10.1145/3298689.3347053.

ERTUGRUL, D. C.; ELCI, A. A survey on semanticized and personalized health recommender systems. **Expert Systems**, v. 37, n. 4, 2020. Available at: https://doi.org/10.1111/exsy.12519.

ESPÍN, V.; HURTADO, M. V.; NOGUERA, M. Nutrition for elder care: A nutritional semantic recommender system for the elderly. **Expert Systems**, v. 33, n. 2, p. 201–210, 2016. Available at: https://doi.org/10.1111/exsy.12143.

ETEMADI, M. *et al.* A systematic review of healthcare recommender systems: Open issues, challenges, and techniques. **Expert Systems with Applications**, v. 213, 2023. Available at: https://doi.org/10.1016/j.eswa.2022.118823.

EVANS, J. S. B. T.; STANOVICH, K. E. Dual-process theories of higher cognition: Advancing the debate. **Perspectives on Psychological Science**, v. 8, n. 3, p. 223–241, 2013. Available at: https://doi.org/10.1177/1745691612460685.

FANG, Y.-H. *et al.* **A step-by-step introduction to the implementation of automatic differentiation**. 2024. Available at: https://arxiv.org/abs/2402.16020.

FERRETTO, L. R.; CERVI, C. R.; MARCHI, A. C. B. de. Recommender systems in mobile apps for health a systematic review. *In*: **12th Iberian Conference on Information Systems and Technologies (CISTI)**. IEEE, 2017. p. 1–6. Available at: http://doi.org/10.23919/CISTI.2017.7975743.

FERRIOLLI, E. *et al.* Assessment of intrinsic capacity in the Brazilian older population and the psychometric properties of the WHO/ICOPE screening tool: A multicenter cohort study protocol. **Geriatrics, Gerontolology and Aging**, v. 18, n. 166, 2024. Available at: http://www.ggaging.com/details/1841.

FOTAKIS, D.; KALAVASIS, A.; PSAROUDAKI, E. Label ranking through nonparametric regression. *In*: CHAUDHURI, K. *et al.* (ed.). **Proceedings of the 39th International Conference on Machine Learning**. PMLR, 2022. (Proceedings of Machine Learning Research, v. 162), p. 6622–6659. Available at: https://proceedings.mlr.press/v162/fotakis22a.html.

FRIKHA, G. *et al.* Leveraging service supply dynamics in senselife: Building an explainable recommender system for tailored frailty prevention. *In*: **21st International Conference on Computer Systems and Applications (AICCSA)**. IEEE, 2024. p. 1–6. Available at: https://doi.org/10.1109/AICCSA63423.2024.10912596.

FÜRNKRANZ, J.; HÜLLERMEIER, E. Preference learning: An introduction. *In*: FÜRNKRANZ, J.; HÜLLERMEIER, E. (ed.). **Preference Learning**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 1–17. ISBN 978-3-642-14125-6. Available at: https://doi.org/10.1007/978-3-642-14125-6_1.

GANNOD, G. *et al.* A machine learning recommender system to tailor preference assessments to enhance person-centered care among nursing home residents.

Gerontologist, Oxford Academic Press, UK, v. 59, n. 1, p. 167–176, 2019. Available at: https://doi.org/10.1093/geront/gny056.

GARRARD, J. *et al.* Comprehensive geriatric assessment in primary care: a systematic review. **Aging Clinical and Experimental Research**, Springer, Germany, v. 32, n. 2, p. 197–205, 2020. Available at: https://doi.org/10.1007/s40520-019-01183-w.

GIERE, R. N. **Scientific Perspectivism**. Chicago: University of Chicago Press, 2006. ISBN 9780226292144. Available at: https://press.uchicago.edu/ucp/books/book/chicago/S/bo4094708.html.

GILBERT, S. The EU passes the AI Act and its implications for digital medicine are unclear. **npj Digital Medicine**, Nature Publishing Group UK London, v. 7, n. 1, p. 135, 2024. Available at: https://doi.org/10.1038/s41746-024-01116-6.

GILLIES, S. *et al.* **Shapely**. Zenodo, 2022. Software for manipulation and analysis of geometric objects in the Cartesian plane. Available at: https://doi.org/10.5281/zenodo.7366742.

GIONIS, A. *et al.* Algorithms for discovering bucket orders from data. *In*: **Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2006. (KDD '06), p. 561–566. ISBN 1595933395. Available at: https://doi.org/10.1145/1150402.1150468.

GOBBENS, R. J. *et al.* The Tilburg Frailty Indicator: Psychometric properties. **Journal of the American Medical Directors Association**, v. 11, n. 5, p. 344–355, 2010. ISSN 1525-8610. Available at: https://doi.org/10.1016/j.jamda.2009.11.003.

GOBBENS, R. J. J. *et al.* The predictive validity of the Tilburg Frailty Indicator: Disability, health care utilization, and quality of life in a population at risk. **The Gerontologist**, v. 52, n. 5, p. 619–631, 01 2012. ISSN 0016-9013. Available at: https://doi.org/10.1093/geront/gnr135.

GORZONI, M. L. Geriatria: Medicina do século XXI? **Medicina (Ribeirão Preto)**, v. 50, n. 3, p. 144–9, 2017. Available at: https://doi.org/10.11606/issn.2176-7262.v50i3p144-149.

HAITSMA, K. V. *et al.* The preferences for everyday living inventory: Scale development and description of psychosocial preferences responses in community-dwelling elders. **Gerontologist**, Oxford Academic Press, UK, v. 53, n. 4, p. 582–595, 2013. Available at: https://doi.org/10.1093/geront/gns102.

HANCOCK, G. R.; AN, J. A closed-form alternative for estimating omega reliability under unidimensionality. **Measurement: Interdisciplinary Research and Perspectives**, Routledge, v. 18, n. 1, p. 1–14, 2020. Available at: https://doi.org/10.1080/15366367.2019.1656049.

HAUSMAN, D. M. **Valuing health: Well-being, freedom, and suffering**. New York, NY, USA: Oxford University Press, 2015. ISBN 978-0-19-0233118-1. Available at: https://global.oup.com/academic/product/valuing-health-9780190233181. Access at: 9/27/2025.

HAYES, A. F.; COUTTS, J. J. Use omega rather than Cronbach's alpha for estimating reliability. but... **Communication Methods and Measures**, Routledge, v. 14, n. 1, p. 1–24, 2020. Available at: https://doi.org/10.1080/19312458.2020.1718629.

HERRERA, F. *et al.* Case studies and metrics. *In*: **Multilabel Classification : Problem Analysis, Metrics and Techniques**. Cham: Springer International Publishing, 2016. p. 33–63. ISBN 978-3-319-41111-8. Available at: https://doi.org/10.1007/978-3-319-41111-8_3.

HOFMANN, T.; SCHÖLKOPF, B.; SMOLA, A. J. Kernel methods in machine learning. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 36, n. 3, p. 1171 – 1220, 2008. Available at: https://doi.org/10.1214/009053607000000677.

HOLT, J. C. Kappenburg-ten. **A comparison between factor analysis and item response theory modeling in scale analysis**. 2014. Tese (Doutorado) — University of Groningen, São Paulo, Brazil, 2014. Available at: https://research.rug.nl/en/publications/a-comparison-between-factor-analysis-and-item-response-theory-mod.

HORS-FRAILE, S. *et al.* Analyzing recommender systems for health promotion using a multidisciplinary taxonomy: A scoping review. **International Journal of Medical Informatics**, v. 114, p. 143 – 155, 2018. Available at: https://doi.org/10.1016/j.ijmedinf.2017.12.018.

HOYNINGEN-HUENE, P. Systematicity: The nature of science. **Philosophia**, v. 36, n. 2, p. 167–180, June 2008. ISSN 1574-9274. Available at: https://doi.org/10.1007/s11406-007-9100-x.

HSIAO, F.-Y.; CHEN, L.-K. Intrinsic capacity assessment works—let's move on actions. **The Lancet Healthy Longevity**, Elsevier, v. 5, n. 7, p. e448–e449, 2024. Available at: https://doi.org/10.1016/S2666-7568(24)00110-7.

HUANG, X.; PALAOAG, T. D. Implementation of intelligent elderly care system based on cloud platform and nnb algorithm. *In*: **2nd International Conference on Control, Electronics and Computer Technology (ICCECT)**. IEEE, 2024. p. 841–846. Available at: https://doi.org/10.1109/ICCECT60629.2024.10546115.

JANSEN, C. *et al.* Statistical multicriteria benchmarking via the GSD-front. *In*: GLOBERSON, A. *et al.* (ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2024. v. 37, p. 98143–98179. Available at: https://proceedings.neurips.cc/paper_files/paper/2024/file/b1f140eeee243db24e9e006481b91cf1-Paper-Conference.pdf.

JIA, B.-B.; LIU, J.-Y.; ZHANG, M.-L. Towards exploiting linear regression for multi-class/multi-label classification: An empirical analysis. **International Journal of Machine Learning and Cybernetics**, v. 15, n. 9, p. 3671–3700, sep 2024. ISSN 1868-808X. Available at: https://doi.org/10.1007/s13042-024-02114-6.

JIANG, F. Personalized recommendation method of health education resources in home-based aged care based on similarity correlation. *In*: **Proceedings of the 2025 International Conference on Digital Education and Information Technology**. New York, NY, USA: ACM, 2025. (DEIT '25), p. 69–73. ISBN 9798400713248. Available at: https://doi.org/10.1145/3732299.3732313.

JIMÉNEZ, J. C. A. **Design of efficient and scalable algorithms for label ranking problems**. 2023. Tese (Doutorado) — Universidad de Castilla-La Mancha, Castilla–La Mancha, Spain, 2023. Available at: http://hdl.handle.net/10578/31844.

JUNG, H.-W. Visualizing domains of comprehensive geriatric assessments to grasp frailty spectrum in older adults with a radar chart. **Annals of Geriatric Medicine and Research**, Korean Geriatrics Society, South Korea, v. 24, n. 1, p. 55–56, 2020. Available at: https://doi.org/10.4235/agmr.20.0013.

KAHNEMAN, D.; KLEIN, G. Conditions for intuitive expertise: A failure to disagree. **The American Psychologist**, v. 64, n. 6, p. 515–526, Sep 2009. ISSN 0003-066X, 1935-990X.

KAMRAN, M.; JAVED, A. A survey of recommender systems and their application in healthcare. **Technical Journal of University of Engineering & Technology Taxila**, v. 20, n. 4, 2015. Available at: https://tj.uettaxila.edu.pk/older-issues/2015/No4/15.A%20Survey%20of%20Recommender%20Systems%20and%20Their%20Application%20in%20Healthcare.pdf.

KANEDA, T.; LEE, M.; POLLARD, K. **SCL/PRB index of well-being in older populations**. Stanford, CA, US, 2011. Available at: https://tinyurl.com/kaneda2011scl. Access at: 10/10/2020.

KATZ, S. *et al.* Studies of illness in the aged: The index of ADL: A standardized measure of biological and psychosocial function. **JAMA**, v. 185, n. 12, p. 914–919, 09 1963. ISSN 0098-7484. Available at: https://doi.org/10.1001/jama.1963.03060120024016.

KAVULURU, R.; RIOS, A.; LU, Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. **Artificial Intelligence in Medicine**, v. 65, n. 2, p. 155–166, 2015. ISSN 0933-3657. Intelligent healthcare informatics in big data era. Available at: https://doi.org/10.1016/j.artmed.2015.04.007.

KNIJNENBURG, B. P.; WILLEMSEN, M. C. Evaluating recommender systems with user experiments. *In*: RICCI, F.; ROKACH, L.; SHAPIRA, B. (ed.). **Recommender Systems Handbook**. 2nd. ed. Boston, USA: Springer, 2015. cap. 9, p. 309–352. ISBN 978-1-4899-7637-6. Available at: https://doi.org/10.1007/978-1-4899-7637-6_9.

KOLAKOWSKI, M. *et al.* CAREUP: An integrated care platform with intrinsic capacity monitoring and prediction capabilities. **Sensors**, v. 25, n. 3, 2025. Available at: https://doi.org/10.3390/s25030916.

KRABBE, P. **The measurement of health and health status: Concepts, methods and applications from a multidisciplinary perspective**. Academic Press, 2016. ISBN 978-0-12-801504-9. Available at: https://doi.org/10.1016/C2013-0-19200-8. Access at: 9/27/2025.

LEDLEY, R. S. Using electronic computers in medical diagnosis. **IRE Transactions on Medical Electronics**, ME-7, n. 4, p. 274–280, 1960. Available at: https://doi.org/10.1109/IRET-ME.1960.5008080.

LEE, Y.; LIM, W. Shoelace formula: Connecting the area of a polygon with vector cross product. **Mathematics Teacher**, National Council of Teachers of Mathematics, USA, v. 110, n. 8, p. 631–636, 2017. ISSN 2330-0582. Available at: https://doi.org/10.5951/mathteacher.110.8.0631.

LEGG, C. What is a logical diagram? *In*: MOKTEFI, A.; SHIN, S.-J. (ed.). **Visual Reasoning with Diagrams**. Basel: Springer Basel, 2013. p. 1–18. ISBN 978-3-0348-0600-8. Available at: https://doi.org/10.1007/978-3-0348-0600-8_1.

LIMA, A. P. de *et al.* **An Interpretable Recommendation Model for Psychometric Data, With an Application to Gerontological Primary Care**. 2026. Available at: https://arxiv.org/abs/2601.19824.

LIMA, A. P. de *et al.* An interpretable recommendation model for gerontological care. *In*: **Proceedings of the 15th ACM Conference on Recommender Systems**. New York, NY, USA: ACM, 2021. (RecSys '21), p. 620–626. ISBN 9781450384582. Available at: https://doi.org/10.1145/3460231.3478850.

LIMA-COSTA, M. F. *et al.* The Brazilian Longitudinal Study of Aging (ELSI-Brazil): Objectives and Design. **American Journal of Epidemiology**, v. 187, n. 7, p. 1345–1353, 01 2018. ISSN 0002-9262. Available at: https://doi.org/10.1093/aje/kwx387.

LIPTON, Z. C. The mythos of model interpretability. **Queue**, ACM, v. 16, n. 3, p. 31–57, 2018. Available at: https://doi.org/10.1145/3236386.3241340.

LLORENTE, Á. *et al.* Assessment of cognitive games to improve the quality of life of Parkinson's and Alzheimer's patients. **Digital Health**, v. 10, 2024. Available at: https://doi.org/10.1177/20552076241254733.

LORENZI, L. *et al.* Digital engagement and quality of life of participants at a University of the Third Age. **Gerontechnology**, v. 21, 2022. Available at: https://doi.org/10.4017/gt.2022.21.s.567.opp4.

LUO, G.; TANG, C.; THOMAS, S. B. Intelligent personal health record: Experience and open issues. *In*: SIGHIT. **Proceedings of the 1st ACM International Health Informatics Symposium**. New York, USA: ACM, 2010. (IHI), p. 326–335. Available at: https://doi.org/10.1145/1882992.1883039.

MADDIO, S.; PELOSI, G.; SELLERI, S. A brief history of the Smith chart. *In*: **2023 8th IEEE History of Electrotechnology Conference (HISTELCON)**. IEEE, 2023. p. 39–41. Available at: https://doi.org/10.1109/HISTELCON56357.2023.10365794.

MADJAROV, G. *et al.* An extensive experimental comparison of methods for multi-label learning. **Pattern Recognition**, v. 45, n. 9, p. 3084–3104, 2012. ISSN 0031-3203. Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011). Available at: https://doi.org/10.1016/j.patcog.2012.03.004.

MAMALAKIS, M. *et al.* The explanation necessity for Healthcare AI. **arXiv e-prints**, May 2024. Available at: https://doi.org/10.48550/arXiv.2406.00216.

MARTÍN, C. *et al.* DigiHEALTH: Suite of digital solutions for long-term healthy and active aging. **International Journal of Environmental Research and Public Health**, v. 20, n. 13, 2023. Available at: https://doi.org/10.3390/ijerph20136200.

MARTIN, E. M.; COSTA, A.; CAZORLA, M. PHAROS 2.0—A PHysical Assistant RObot System Improved. **Sensors**, Multidisciplinary Digital Publishing Institute, Switzerland, v. 19, n. 20, p. 4531, 2019. ISSN 1424-8220. Available at: https://doi.org/10.3390/s19204531.

MCNEISH, D. Thanks coefficient alpha, we'll take it from here. **Psychological Methods**, American Psychological Association, Washington, DC, USA, v. 23, n. 3, p. 412–433, 2018. ISSN 1082-989X. Available at: https://doi.org/10.1037/met0000144.

MCNEISH, D. Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. **Behavior Research Methods**, v. 55, n. 8, p. 4269–4290, 2023. ISSN 1554-3528. Available at: https://doi.org/10.3758/s13428-022-02016-x.

MCNEISH, D. Practical implications of sum scores being psychometrics' greatest accomplishment. **Psychometrika**, v. 89, n. 4, p. 1148–1169, 2024. Available at: https://doi.org/10.1007/s11336-024-09988-z.

MCNEISH, D.; WOLF, M. G. Thinking twice about sum scores. **Behavior Research Methods**, v. 52, n. 6, p. 2287–2305, 2020. ISSN 1554-3528. Available at: https://doi.org/10.3758/s13428-020-01398-0.

MELO, R. C. de; SILVA, T. B. L. da; CACHIONI, M. Desafios da formação em gerontologia. **Revista Kairós-Gerontologia**, v. 18, p. 123–147, 2015. Available at: https://revistas.pucsp.br/index.php/kairos/article/view/27261/19297.

MICHELL, J. **Measurement in psychology: A critical history of a methodological concept**. Cambridge, UK: Cambridge University Press, 1999. ISBN 0-521-62120-8. Available at: https://www.cambridgebookshop.co.uk/products/measurement-in-psychology.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. **Artificial Intelligence**, Elsevier, v. 267, p. 1–38, 2019. Available at: https://doi.org/10.1016/j.artint.2018.07.007.

MINAKATA, M. *et al.* Safe walking route recommender based on fall risk calculation using a digital human model on a 3d map. **IEEE Access**, v. 10, p. 8424 – 8433, 2022. Available at: https://doi.org/10.1109/ACCESS.2022.3143322.

MINEMAWARI, Y.; KATO, T. The radar chart method and its analysis as a comprehensive geriatric assessment system for elderly disabled patients. **Japanese Geriatrics Society**, Japan Journal of Geriatrics, Japan, v. 36, n. 3, p. 206–212, 1999. ISSN 0300-9173. Available at: https://doi.org/10.3143/geriatrics.36.206.

MISHRA, P.; BUSETTY, S. M.; GUDLA, S. K. Enhanced activity recognition of the IoT smart home users through cluster analysis. *In*: **6th World Forum on Internet of Things (WF-IoT)**. virtual: IEEE, 2020. p. 1–6.

MULLAN, F. Wrestling with variation: An interview with Jack Wennberg. **Health Affairs**, v. 23, n. Supplement 2, p. VAR–73–VAR–80, 2004. Available at: https://doi.org/10.1377/hlthaff.var.73.

MURAKAMI, K. *et al.* Information recommendation system for the care prevention using a communication robot. *In*: **Proceedings of Society of Instrument and Control Engineers (SICE) Annual Conference 2010**. IEEE, 2010. p. 388–389. Available at: https://ieeexplore.ieee.org/document/5603071.

NEEDHAM, T. **Visual Complex Analysis**. UK: Oxford University Press, 1997/2012. ISBN 978-0-19-853446-4.

NUNES, I.; JANNACH, D. A systematic review and taxonomy of explanations in decision support and recommender systems. **User Modeling and User-Adapted Interaction**, Springer, Netherlands, v. 27, n. 3-5, p. 393–444, 2017. Available at: https://doi.org/10.1007/s11257-017-9195-0.

O'DONNELL, C. A. Variation in GP referral rates: What can we learn from the literature? **Family Practice**, v. 17, n. 6, p. 462–471, 12 2000. ISSN 0263-2136. Available at: https://doi.org/10.1093/fampra/17.6.462.

OLIVA-FELIPE, L. *et al.* Health recommender system design in the context of CAREGIVERSPRO-MMD Project. *In*: **Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference**. New York, USA: ACM, 2018. (PETRA), p. 462–469. Available at: https://doi.org/10.1145/3197768.3201558.

ORLEY, J. *et al.* **WHOQOL-BREF: Introduction, administration, scoring and generic version of the assessment: field trial version**. Geneva, 1996. Available at: https://www.who.int/publications/i/item/WHOQOL-BREF. Access at: 7/27/2024.

ORTE, S. *et al.* Dynamic decision support system for personalised coaching to support active ageing. *In*: **AI*AAL.it 2018 - Fourth Italian Workshop on Artificial Intelligence for Ambient Assisted Living**. CEUR Workshop Proceedings, 2018. v. 2333, p. 16–36. Available at: http://ceur-ws.org/Vol-2333/paper2.pdf.

OTTO, M. P. **Scalable and interpretable kernel methods based on random Fourier features**. 2023. Dissertação (Mestrado) — Estatística Interinstitucional do ICMC e UFSCar, São Carlos, Brazil, 2023. Available at: https://doi.org/10.11606/D.104.2023.tde-02052023-084042.

PHILIPPI, C. L. **On the Challenges of Measurement in the Human Sciences**. 2023. Tese (Doutorado) — Apollo - University of Cambridge Repository, 2023. Available at: https://www.repository.cam.ac.uk/handle/1810/358705.

PINCAY, J.; TERÁN, L.; PORTMANN, E. Health recommender systems: A state-of-the-art review. *In*: **Proceedings of the 6th International Conference on eDemocracy & eGovernment**. New York, USA: IEEE, 2019. (ICEDEG), p. 47–55. Available at: https://doi.org/10.1109/ICEDEG.2019.8734362.

PONCE, V. *et al.* QueFaire: Context-aware in-person social activity recommendation system for active aging. *In*: **Inclusive Smart Cities and e-Health**. Netherlands: Springer, 2015, (Lecture Notes in Computer Science, v. 9102). p. 64–75. ISBN 978-3-319-19312-0. Available at: https://doi.org/10.1007/978-3-319-19312-0_6.

PRASETYO, R. F.; BAIZAL, Z. K. A. Ontology-based healthy food recommendation for the elderly. *In*: **International Conference on Communication, Networks and Satellite (COMNETSAT)**. IEEE, 2024. p. 76–81. Available at: https://doi.org/10.1109/COMNETSAT63286.2024.10862504.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Recommender systems: Introduction and challenges. *In*: RICCI, F.; ROKACH, L.; SHAPIRA, B. (ed.). **Recommender**

Systems Handbook. 2nd. ed. Boston, USA: Springer, 2015. cap. 1, p. 1–34. ISBN 978-1-4899-7637-6. Available at: https://doi.org/10.1007/978-1-4899-7637-6_1.

RINCON, J. A. *et al.* A new emotional robot assistant that facilitates human interaction and persuasion. **Knowledge and Information Systems**, Springer, UK, v. 60, n. 1, p. 363–383, 2019. Available at: https://doi.org/10.1007/s10115-018-1231-9.

RIST, T. *et al.* CARE - Extending a digital picture frame with a recommender mode to enhance well-being of elderly people. *In*: **Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare**. New York, USA: IEEE, 2015. (PervasiveHealth), p. 112–120. Available at: https://doi.org/10.4108/icst.pervasivehealth.2015.259255.

RIVOLLI, A.; PARKER, L. C.; CARVALHO, A. C. P. L. F. de. Food truck recommendation using multi-label classification. *In*: OLIVEIRA, E. *et al.* (ed.). **Progress in Artificial Intelligence**. Cham: Springer International Publishing, 2017. p. 585–596. ISBN 978-3-319-65340-2.

RIVOLLI, A. *et al.* An empirical analysis of binary transformation strategies and base algorithms for multi-label learning. **Machine Learning**, v. 109, n. 8, p. 1509–1563, 2020. ISSN 1573-0565. Available at: https://doi.org/10.1007/s10994-020-05879-3.

ROBINSON, J.; APPIAH, K.; YOUSAF, R. Improving the well-being of older people by reducing their energy consumption through energy-aware systems. *In*: **Proceedings of the 9th International Conference on eHealth, Telemedicine, and Social Medicine**. International Academy, Research and Industry Association (IARIA), 2017. (eTELEMED). ISBN 9781612085401. Available at: http://irep.ntu.ac.uk/id/eprint/30483.

ROMÁN-VILLARÁN, E. *et al.* A personalized ontology-based decision support system for complex chronic patients: Retrospective observational study. **JMIR Formative Research**, v. 6, n. 8, 2022. Available at: https://doi.org/10.2196/27990.

ROSS, T. J. Fuzzy classification. *In*: ROSS, TIMOTHY J. **Fuzzy Logic with Engineering Applications**. 3rd. ed. John Wiley & Sons, Ltd, 2010. cap. 10, p. 332–368. ISBN 9781119994374. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119994374.ch10.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. **Nature Machine Intelligence**, v. 1, n. 5, p. 206–215, 2019. Available at: https://doi.org/10.1038/s42256-019-0048-x.

SAID, A. *et al.* Fifth International Workshop on Health Recommender Systems (HealthRecSys). *In*: SIGCHI. **Proceedings of the 14th ACM Conference on Recommender Systems**. New York, USA: ACM, 2020. (RecSys), p. 611–612. Available at: https://doi.org/10.1145/3383313.3411540.

SAKET, B.; ENDERT, A.; DEMIRALP, Ç. Task-based effectiveness of basic visualizations. **IEEE Transactions on Visualization and Computer Graphics**, v. 25, n. 7, p. 2505–2512, 2019. Available at: https://doi.org/10.1109/TVCG.2018.2829750.

SASAKI, W.; TAKAMA, Y. Walking route recommender system considering SAW criteria. *In*: **Conference on Technologies and Applications of Artificial Intelligence**. IEEE, 2013. p. 246–251. Available at: https://doi.org/10.1109/TAAI.2013.56.

SCHÖLKOPF, B.; TSUDA, K.; VERT, J.-P. **Kernel Methods in Computational Biology**. The MIT Press, 2004. ISBN 9780262256926. Available at: https://doi.org/10.7551/mitpress/4057.001.0001.

SEZGIN, E.; ÖZKAN, S. A systematic literature review on Health Recommender Systems. *In*: **Proceedings of the E-Health and Bioengineering Conference**. New York, USA: IEEE, 2013. (EHB), p. 1–4. ISBN 978-1-4799-2372-4. Available at: http://doi.org/10.1109/EHB.2013.6707249.

SHASHAR, S. *et al.* Unravelling the determinants of medical practice variation in referrals among primary care physicians: insights from a retrospective cohort study in Southern Israel. **BMJ Open**, British Medical Journal Publishing Group, v. 13, n. 8, 2023. ISSN 2044-6055. Available at: https://bmjopen.bmj.com/content/13/8/e072837.

SHENK, D. *et al.* Teaching research in Gerontology: Toward a cumulative model. **Educational Gerontology**, Routledge, v. 27, n. 7, p. 537–556, 2001. Available at: https://doi.org/10.1080/036012701753122884.

SHIBATA, S. *et al.* Development of a health promotion system for the elderly: Committee of Health Evaluation for Elderly Persons Council of Japan AMHTS Institutions. **Journal of Medical Systems**, Springer, USA, v. 22, n. 1, p. 43–49, 1998. ISSN 1573-689X. Available at: http://doi.org/10.1023/A:1022654406018.

SHINDE, K. *et al.* An IoT-based context-aware recommender system to improve the quality of life of elderly people. *In*: **11th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)**. IEEE, 2021. v. 1, p. 202–206. Available at: https://doi.org/10.1109/IDAACS53288.2021.9660935.

SIJTSMA, K.; ELLIS, J. L.; BORSBOOM, D. Recognize the value of the sum score, psychometrics' greatest accomplishment. **Psychometrika**, v. 89, n. 1, p. 84–117, 2024. Available at: https://doi.org/10.1007/s11336-024-09964-7.

SIJTSMA, K.; ELLIS, J. L.; BORSBOOM, D. Rejoinder to mcneish and mislevy: What does psychological measurement require? **Psychometrika**, v. 89, n. 4, p. 1175–1185, 2024. Available at: https://doi.org/10.1007/s11336-024-10004-7.

SILVA, P. A. B. *et al.* Cut-off point for WHOQOL-BREF as a measure of quality of life of older adults. **Revista de Saúde Pública**, SciELO Brasil, v. 48, p. 390–397, 2014. Available at: https://doi.org/10.1590/S0034-8910.2014048004912.

SILVA, T. *et al.* +TV4E - Delivering information about social services for seniors throughout TV. *In*: INSTICC. **Opportunities and Challenges for European Projects - Volume 1: EPS Portugal 2017/2018**. Portugal: SciTePress, 2017. p. 133–149. ISBN 978-989-758-361-2. Available at: https://doi.org/10.5220/0008862501330149.

SITPAROOPAN, T. *et al.* Home Bridge - Smart elderly care system. *In*: **2nd International Informatics and Software Engineering Conference (IISEC)**. IEEE, 2021. p. 1–5. Available at: https://doi.org/10.1109/IISEC54230.2021.9672411.

SKEVINGTON, S. M.; LOTFY, M.; O'CONNELL, K. A. The world health organization's WHOQOL-BREF quality of life assessment: Psychometric properties and results of the international field trial. A report from the WHOQOL group. **Quality**

of Life Research, v. 13, n. 2, p. 299–310, 2004. ISSN 1573-2649. Available at: https://doi.org/10.1023/B:QURE.0000018486.91360.00.

STEINWART, I. Support vector machines are universally consistent. **Journal of Complexity**, v. 18, n. 3, p. 768–791, 2002. ISSN 0885-064X. Available at: https://doi.org/10.1006/jcom.2002.0642.

STEMLER, S. E. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. **Practical Assessment, Research, and Evaluation**, University of Maryland, USA, v. 9, n. 1, p. 4, 2004. ISSN 1531-7714. Available at: https://doi.org/10.7275/96jp-xz07.

STILLER, C.; ROSS, F.; AMENT, C. Demographic recommendations for WEITBLICK, an assistance system for elderly. *In*: **10th International Symposium on Communications and Information Technologies**. New York, USA: IEEE, 2010. p. 406–411. Available at: https://doi.org/10.1109/ISCIT.2010.5664874.

SU, J. *et al.* **Do recommender systems function in the health domain: a system review**. Ithaca, USA: Cornell University, 2020. Available at: https://arxiv.org/abs/2007.13058.

TAL, E. Measurement in Science. *In*: ZALTA, E. N. (ed.). **The Stanford Encyclopedia of Philosophy**. Fall 2020. Stanford, CA, US: Metaphysics Research Lab, Stanford University, 2020. Available at: https://plato.stanford.edu/archives/fall2020/entries/measurement-science/.

TANIMOTO, T. IBM Type 704 medical diagnosis program. **IRE Transactions on Medical Electronics**, ME-7, n. 4, p. 280–283, 1960. Available at: https://doi.org/10.1109/IRET-ME.1960.5008081.

TAVASSOLI, N. *et al.* Implementation of the WHO Integrated Care for Older People (ICOPE) programme in clinical practice: A prospective study. **The Lancet Healthy Longevity**, v. 3, n. 6, p. e394–e404, 2022. ISSN 2666-7568. Available at: https://doi.org/10.1016/S2666-7568(22)00097-6.

TEBET, R. **Estatuto do idoso - Lei no 10.741/2003**. Brasília, 2006. Available at: https://www2.senado.leg.br/bdsf/item/id/530232. Access at: 1/9/2021.

THAKUR, N.; HAN, C. Y. A context-driven complex activity framework for smart home. *In*: **9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)**. IEEE, 2018. p. 801–806. Available at: https://doi.org/10.1109/IEMCON.2018.8615079.

THISSEN, D. *et al.* An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. **Applied Psychological Measurement**, v. 7, n. 2, p. 211–226, 1983. Available at: https://doi.org/10.1177/014662168300700209.

TINTAREV, N.; MASTHOFF, J. Beyond explaining single item recommendations. *In*: RICCI, F.; ROKACH, L.; SHAPIRA, B. (ed.). **Recommender Systems Handbook**. New York, NY: Springer US, 2022. p. 711–756. ISBN 978-1-0716-2197-4. Available at: https://doi.org/10.1007/978-1-0716-2197-4_19.

TRACTINSKY, N. The usability construct: A dead end? **Human-Computer Interaction**, Taylor & Francis, USA, v. 33, n. 2, p. 131–177, 2018. Available at: https://doi.org/10.1080/07370024.2017.1298038.

TRAN, T. N. T. *et al.* Recommender systems in the healthcare domain: State-of-the-art and research issues. **Journal of Intelligent Information Systems**, v. 57, n. 1, p. 171 – 201, 2021. Available at: https://doi.org/10.1007/s10844-020-00633-6.

TSOUMAKAS, G.; KATAKIS, I. Multi-label classification: An overview. **International Journal of Data Warehousing and Mining (IJDWM)**, IGI Global Scientific Publishing, v. 3, n. 3, p. 1–13, 2007.

UNWIN, A.; KLEINMAN, K. The iris data set: In search of the source of virginica. **Significance**, v. 18, n. 6, p. 26–29, 11 2021. ISSN 1740-9705. Available at: https://doi.org/10.1111/1740-9713.01589.

VARKEY, B. Principles of clinical ethics and their application to practice. **Medical Principles and Practice**, v. 30, n. 1, p. 17–28, 06 2020. ISSN 1011-7571. Available at: https://doi.org/10.1159/000509119.

VERCELLI, A. *et al.* My-Active and Healthy Ageing (My-AHA): An ICT platform to detect frailty risk and propose intervention. *In*: **25th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)**. IEEE, 2017. p. 1–4. Available at: https://doi.org/10.23919/SOFTCOM.2017.8115505.

VERGANI, C. *et al.* A polar diagram for comprehensive geriatric assessment. **Archives of Gerontology and Geriatrics**, Elsevier, Ireland, v. 38, n. 2, p. 139–144, 2004. ISSN 0167-4943. Available at: https://doi.org/10.1016/j.archger.2003.08.009.

VIEIRA, A. *et al.* A systematic review on recommendation systems applied to chronic diseases. **Intelligent Data Analysis**, v. 27, n. 5, p. 1223 – 1265, 2023. Available at: https://doi.org/10.3233/IDA-220394.

WATSON, D. S. Conceptual challenges for interpretable machine learning. **Synthese**, v. 200, n. 2, p. 65, 2022. ISSN 1573-0964. Available at: https://doi.org/10.1007/s11229-022-03485-5.

WELSH, T.; GORDON, A.; GLADMAN, J. Comprehensive geriatric assessment - a guide for the non-specialist. **International Journal of Clinical Practice**, v. 68, n. 3, p. 290 – 293, 2014. Available at: https://doi.org/10.1111/ijcp.12313.

WIDAMAN, K. F.; REVELLE, W. Thinking thrice about sum scores, and then some more about measurement and analysis. **Behavior Research Methods**, v. 55, n. 2, p. 788–806, 2023. ISSN 1554-3528. Available at: https://doi.org/10.3758/s13428-022-01849-w.

WIDAMAN, K. F.; REVELLE, W. Thinking about sum scores yet again, maybe the last time, we don't know, oh no...: A comment on mcneish (2023). **Educational and Psychological Measurement**, v. 84, n. 4, p. 637–659, 2024. Available at: https://doi.org/10.1177/00131644231205310.

World Health Organization. **Programme on mental health: WHOQOL user manual**. Geneva, 2012. Available at: https://iris.who.int/handle/10665/77932. Access at: 10/6/2025.

World Health Organization. **World report on ageing and health (full report)**. Geneva, 2015. Available at: https://apps.who.int/iris/handle/10665/186463. Access at: 6/30/2024.

World Health Organization. **Integrated Care for Older People: Guidelines on Community-Level Interventions to Manage Declines in Intrinsic Capacity**. Geneva: World Health Organization, 2017. ISBN 978-92-4-155010-9. Available at: https://www.who.int/publications/i/item/9789241550109. Access at: 9/22/2025.

World Health Organization. **High-value referrals: learning from challenges and opportunities of the COVID-19 pandemic. Concept paper**. Copenhagen, 2023. Available at: https://iris.who.int/handle/10665/367955. Access at: 7/13/2024.

WU, Y. *et al.* Interpretable machine learning for personalized medical recommendations: A LIME-based approach. **Diagnostics**, v. 13, n. 16, 2023. Available at: https://doi.org/10.3390/diagnostics13162681.

WU, Y. *et al.* Interpretable medical recommendations based on SHAPs. *In*: **6th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)**. IEEE, 2023. p. 977–983. Available at: https://doi.org/10.1109/PRAI59366.2023.10332022.

YANG, J.; SHI, R.; NI, B. MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis. *In*: **2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)**. IEEE, 2021. p. 191–195. Available at: https://doi.org/10.1109/ISBI48211.2021.9434062.

YU, J.; CHENG, Q.; HUANG, H. Analysis of the weighting exponent in the fcm. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, v. 34, n. 1, p. 634–639, 2004.

YUAN, X.; ZHANG, Y.; ZHOU, Q. Research on recommendation algorithm for short video health content targeting elderly users. *In*: **International Conference on Information Technology, Communication Ecosystem and Management (ITCEM)**. IEEE, 2024. p. 74–79. Available at: https://doi.org/10.1109/ITCEM65710.2024.00022.

ZACHARAKI, E. I. *et al.* Frailsafe: An ict platform for unobtrusive sensing of multi-domain frailty for personalized interventions. **IEEE Journal of Biomedical and Health Informatics**, v. 24, n. 6, p. 1557–1568, 2020. Available at: https://doi.org/10.1109/jbhi.2020.2986918.

ZHOU, Q. Recipe recommendation for the elderly based on collaborative filtering algorithm and neural network. *In*: **2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)**. IEEE, 2024. p. 660–663. Available at: https://doi.org/10.1109/ICSECE61636.2024.10729407.

ZHOU, Y.; QIU, G. Random forest for label ranking. **Expert Systems with Applications**, v. 112, p. 99–109, 2018. ISSN 0957-4174. Available at: https://doi.org/10.1016/j.eswa.2018.06.036.

APPENDIX

APPENDIX A – A LINK BETWEEN SUM-SCORES AND AREA-SCORES

The assessment polygon is a visual element of the Polygrid diagram that corresponds to the polygon that appears in an assessment chart, as seen in Figure 11. Under some assumptions, the area of an assessment polygon has a monotonic relationship with the sum-score of a psychometric instrument. This relationship is important for two reasons:

- It provides a rational justification for the equivalence between the sum-score and the area-score for the purpose of ordering subjects with respect to a latent variable;
- This equivalence warrants identifying the meaning of the measurand with the meaning of area of the assessment polygon, for the purpose of ordering people on the measurand. In other words, the assessment polygon becomes a visual analogue of the measurand.

The remainder of this appendix is organised as follows. Section A.1 revisits the congeneric factor model seen in Section 2.3 to show how its sum-scores are used to order subjects with respect to its latent variable. Section A.2 introduces the notion of area-score and its pictorial representation, and Section A.3 deduces the equivalence between the sum-score and the area-score for the purpose of ordering people with respect to the latent variable. Section A.4 presents empirical evidence in support of this relationship. As said before, the ideas in this appendix are based on some assumptions, which are made explicit as follows:

(a1) The data collected with the instrument reliably fit a congeneric factor model. This means that the instrument measures a unidimensional latent variable. In this sense, a unidimensional variable is a variable that can only assume real numbers as values, and a latent variable is a random variable that represents an attribute that (presently) cannot be measured directly without substantial measurement error. The congeneric model is the "mechanism" that allows us to access the measurand by its indicators;

(a2) The latent variable tracks an attribute that is framed as a capacity rather than a deficit. In this sense, capacity denotes "an ability that people usually have", and deficit denotes "an ability that people are expected to have but miss". Since the variable admits only unidimensional scalars as numerical representation (as stated in a1), it seems a sensible choice to represent it by non-negative real numbers;

(a3) The correlation between the latent variable and each of its indicators is positive. This is a consequence of the congeneric model being a reflective measurement model. According to Tractinsky (2018), "values of the measures change together in the same direction" in such models. This assumption implies that the correlation between any pair of indicators of the latent variable is also positive (see Equation 2.2).

A.1 The structure of data collected with a psychometric instrument

Following the notation established in Section 4.1, let $\mathring{X}$ be an (m, d)-matrix with the scores obtained by m individuals who were evaluated using a psychometric instrument with d domains, with $d > 2$. Thus, $\mathring{x}_{ik}$ denotes the score obtained by the i-th individual on the k-th domain surveyed by that instrument. As said before (a1), we assume the data was collected with an instrument whose subscales were validated against a congeneric factor model. Based on this, I consider valid each of the following statements:

- Each subject is associated with some level η_i of the measurand (i.e., the attribute that the latent variable tracks in people), which cannot be directly observed;
- The measurand is assumed to exert causal power on the cognitive processes a person relies on when he or she responds to the items in a questionnaire. Of course, the response a subject gives to an item is an observable variable. Formally, $\mathring{x}_{ik} = f_k(\eta_i)$. This causal account of the link between a latent variable and its indicators has been defended by Borsboom, Mellenbergh and Heerden (2003) in some settings[1];
- The relationship between a latent variable η_i and the response given to each of its respective items ($\mathring{x}_{ik}$) is linear. Moreover, responses given to distinct items may be more or less influenced by the latent variable. Thus $\mathring{x}_{ik} = f_k(\eta_i) = \lambda_k \eta_i$, where λ_k represents the direct effect of η_i on $\mathring{x}_{ik}$;
- Responses given to items are subject to noise. Any deviations from the expected linear response $\lambda_k \eta_i$ are accounted for as measurement error (ϵ_{ik}). Thus, we arrive at the standard measurement equation formulated at the individual level:

$$\mathring{x}_{ik} = \lambda_k \eta_i + \epsilon_{ik} \,. \tag{A.1}$$

The ability to order subjects according to their position on a latent variable is a highly valued property of psychometric instruments. This ability is useful in gerontological studies that aim to stratify a target population with respect to some risk (Silva *et al.*, 2014; Andrade *et al.*, 2020; Aliberti *et al.*, 2022), which allows proper care to be provided to groups according to their needs. For instance, in Figure 34, how should Participants 030

[1] The authors say that "in a standard measurement model, the causal ingredient of realism can be defended in a between-subjects sense, but not in a within-subjects sense ... To substantiate causal conclusions at the individual level, one must investigate patterns of covariation at the individual level, that is, one must fit within-subject latent variable models to repeated measurements [of the same individuals]." Thus, the authors do not extend this causal account to congeneric models. In this thesis, however, I rely on the tradition in recommender systems, which does not stand on covariance models, and hope any practical issues emerging from this violation can be detected by empirically testing the fit of theory and data in Section A.4.

and 089 be ordered according to their assessments of quality of life? In other words, which participant has a "higher" quality of life? Essentially, this task has two common approaches. The first, more common in the factor analytic tradition, is to sum up the products of all item responses by their respective factor loadings to obtain a weighted score $\hat{\eta}_i = \sum_k \lambda_k \mathring{x}_{ik}$ (see Eq. 2.6). This practice suggests that, since ϵ_{ik} remains unknown after a measurement is made, it is assumed to be negligible. This additional assumption reduces the standard measurement equation in A.1 to $\mathring{x}_{ik} = \lambda_k \eta_i$, and makes $\hat{\eta}_i = \sum_k \lambda_k(\lambda_k \eta_i) = \eta_i \sum_k \lambda_k^2 = C\eta_i$, with C being an arbitrary positive constant. Thus, if two subjects A and B obtain weighted scores $\hat{\eta}_a$ and $\hat{\eta}_b$ such that $\hat{\eta}_a > \hat{\eta}_b$, and the measurement error is negligible, then $C\eta_a > C\eta_b \iff \eta_a > \eta_b$ must hold. This means that subject A has a higher position on the latent variable η than subject B. The second approach is what is usually called the sum-score in the literature on psychometrics. It corresponds to a particular case of the previous approach in which factors $\lambda_j = 1$ for all j (see Eq. 2.4). For convenience, we will refer to scores obtained from both approaches simply as sum-scores from now on.

A.2 The area-score and its pictorial representation

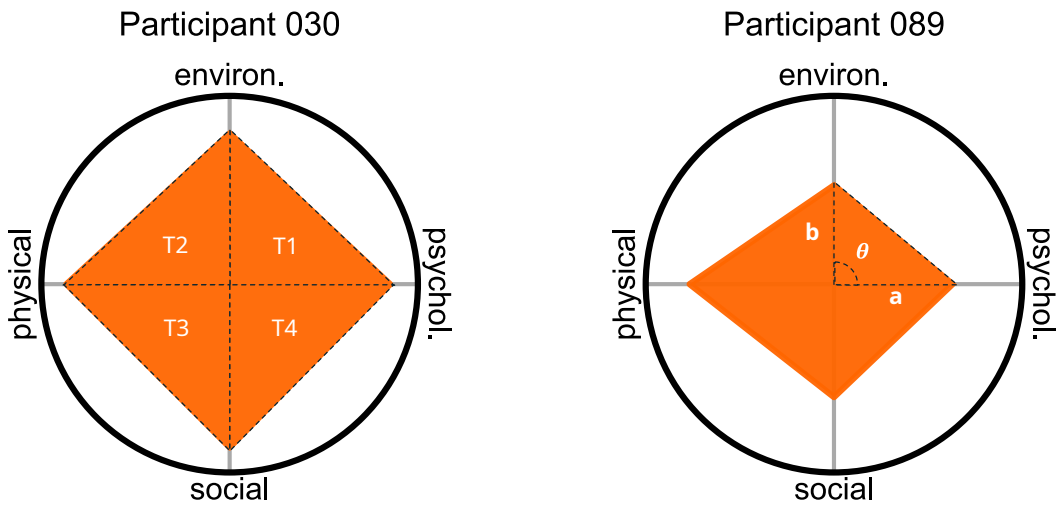
In a Polygrid diagram, the scores a subject obtains when assessed with a psychometric instrument are represented in an assessment chart. Figure 34 shows assessment charts for two subjects from the whoqol dataset. Their scores were described in Table 10. The yellow polygon, which we refer to as an assessment polygon, is the main component of an assessment chart. Its vertices are determined by the subject' scores for each domain of the WHOQOL-BREF: psychological, environmental, physical, and social, as detailed in Section 2.4.1. However, the order in which the vertices are placed on the chart is arbitrary.

We define the **area-score of a subject** (σ_i) as the area of the assessment polygon that represents the scores obtained by a subject. It must be noted that the area of any assessment polygon can be easily evaluated by decomposition. Take the assessment chart of Participant 030 as an example. Its area is the sum of the area of four triangles:

- triangle T1, defined by the scores for the psychological and environmental domains;
- triangle T2, defined by the scores for the environmental and physical domains;
- triangle T3, defined by the scores for the physical and social domains;
- triangle T4, defined by the scores for the social and psychological domains.

The assessment polygon for Participant 089 has been annotated differently. This is to remind us that the area of a triangle with sides a and b, connected by an angle θ, is given by $\frac{\sin\theta}{2}ab$. In this example, with $\theta = \pi/2$, the area is simplified to $\frac{1}{2}ab$. Based on the fact that θ depends only on the number of domains surveyed by the instrument, the area of

Figure 34 – Assessment charts of two subjects evaluated with the WHOQOL-BREF.



Source: The author.

Note: In the assessment chart for Participant 030, the assessment polygon is decomposed into four triangles, T1 to T4. In the assessment polygon for Participant 089, the corresponding triangle T1 appears with two of its sides annotated as a and b, connected by the angle θ.

the triangles covering an assessment polygon[2] can be easily generalised. In other words, since $\theta = 2\pi/d$ for all covering triangles, the area of any such triangle is $\sigma_{ik} := \nu\, x_{ik}\, x_{i,k+1}$, with $\nu = \frac{\sin\theta}{2}$. Clearly, $\nu > 0$ for $d > 2$. Consequently, the area of an assessment polygon is $\sigma_i := \sum_k \sigma_{ik} = \sum_k \nu\, x_{ik}\, x_{i,k+1} = \nu \sum_k x_{ik}\, x_{i,k+1}$, with $(k+1)$ taken modulo d.

A.3 The monotonic relationship between sum-scores and area-scores

We now demonstrate that the area-score induces an ordering of subjects that is equal to the ordering induced on the latent variable by the sum-scores. We show that, under certain assumptions, whenever $\eta_a > \eta_b$ holds, it is the case that a similar condition on the corresponding area-scores $\sigma_a > \sigma_b$ also holds. The assumptions have been stated earlier, but they are formalised here for precision and transparency:

(a1) $\mathring{x}_{ik} = \lambda_k \eta_i + \epsilon_{ik}$ (a structural equation from a congeneric factor model, Eq. 2.1);

(a2) $\eta_i > 0$, for all $i = 0 \ldots m-1$ (the latent variable tracks a capacity of individuals);

(a3) $\lambda_k > 0$, for all $k = 0 \ldots d-1$ (items correlate positively with the latent variable);

(a4) We make an additional assumption: the measurement error is negligible, so $\epsilon_{ik} \approx 0$. This ideal condition is approached as the instrument's reliability increases. Here, we assume that the McDonald's ω from Equation 2.3 is the measure of reliability.

[2] Recall from Section 4.2.1 that the assessment polygon is a polygon whose vertices are scalar multiples of the roots of unity, $\zeta^d = 1$, and they are specified by construction in Algorithm 1.

(a5) The subscales of the instrument have the same range. As seen in Section 2.4.1, the maximum score a person attains in any of the four domains of the WHOQOL-BREF instrument is 5. Recall from Section 4.1 that $\mathring{x}_{ik} \geq 0$ and $x_{ik} = \mathring{x}_{ik}/max\{\mathring{X}_{:k}\}$. If the instrument subscales have the same upper bound $max\{\mathring{X}_{:k}\} = C$ for all $k \in 1 \ldots d$, with $C > 0$, then it is the case that $\mathring{x}_{ak} > \mathring{x}_{bk} \implies \mathring{x}_{ak}/C > \mathring{x}_{bk}/C \implies x_{ak} > x_{bk}$.

Under assumptions (a1) to (a5), whenever there are actual individual differences in the attribute represented by η, namely $\eta_a > \eta_b$, it must be the case that:

$$\eta_a > \eta_b \overset{a3}{\implies} \eta_a\lambda_k > \eta_b\lambda_k \overset{a1,a4}{\implies} \mathring{x}_{ak} > \mathring{x}_{bk} \overset{a5,a2}{\implies} x_{ak} > x_{bk} > 0 \ (\forall k \in 0 \ldots d-1) \quad \text{(A.2)}$$

$$(x_{ak} > x_{bk}) \wedge (x_{a,k+1} > x_{b,k+1}) \implies x_{ak}x_{a,k+1} > x_{bk}x_{b,k+1} \ (\forall k \in 0 \ldots d-1) \quad \text{(A.3)}$$

$$\nu x_{ak}x_{a,k+1} > \nu x_{bk}x_{b,k+1} \implies \sigma_{ak} > \sigma_{bk} \ (\forall k \in 0 \ldots d-1) \quad \text{(A.4)}$$

$$\sigma_{ak} > \sigma_{bk} \implies \sum_k \sigma_{ak} > \sum_k \sigma_{bk} \implies \sigma_a > \sigma_b \quad \text{(A.5)}$$

The statement A.2 speaks about the scores obtained by two subjects for a single domain k. If subject A has a higher position on the latent variable η than subject B, namely $\eta_a > \eta_b$, then a similar relation holds for their scores on the k-th domain, $x_{ak} > x_{bk}$. The statement A.3 expands the previous one to a neighbouring domain: since $x_{ak} > x_{bk} > 0$ holds for the k-th domain, and $x_{a,k+1} > x_{b,k+1} > 0$ holds for the $(k+1)$-th domain, the inequality about the area of two rectangles formed by consecutive scores also holds: $x_{ak}x_{a,k+1} > x_{bk}x_{b,k+1}$. The statement A.4 connects those rectangles to the triangles formed by consecutive scores in an assessment chart. Using the examples in Figure 34 with $k = 0$, the statement $\nu x_{ak}x_{a,k+1} > \nu x_{bk}x_{b,k+1}$ means that the area of the triangle T1 in the assessment polygon of Participant 030 is greater than the area of the corresponding triangle in the assessment polygon of Participant 089. Here, ν has the meaning presented at the end of the previous section, $\nu = \frac{sin\theta}{2}$. Finally, the statement A.5 starts with a fact about the area of triangles formed by consecutive scores, $\sigma_{ak} > \sigma_{bk}$, and concludes with a fact about area-scores. The rightmost side of the statement says that the area-score of subject A, which corresponds to the sum of the area of the triangles that cover its corresponding assessment polygon, $\sigma_a = \sum_k \nu x_{ak}x_{a,k+1}$, is greater than the area-score of subject B, $\sigma_b = \sum_k \nu x_{bk}x_{b,k+1}$.

A.4 Empirical evidence for the reliability of the relationship

The hypothesised relationship between sum-scores and area-scores was tested on our benchmark datasets. We used unit scaled data instead of original data (in $\mathring{X}$) because, unlike WHOQOL-BREF, the AMPIAB and ELSIO1 instruments have domain scales whose range vary greatly, which violates the assumption (a5). The test was designed as follows: (a) for each pair of assessments $(x_a, x_b) \in X^2$, we compute their unit weighted scaled sum-scores $(s_a, s_b) := (\sum_k x_{ak}, \sum_k x_{bk})$; (b) If $(s_a = s_b)$, the pair is discarded from the analysis. Otherwise, the respective area-scores are computed using the same

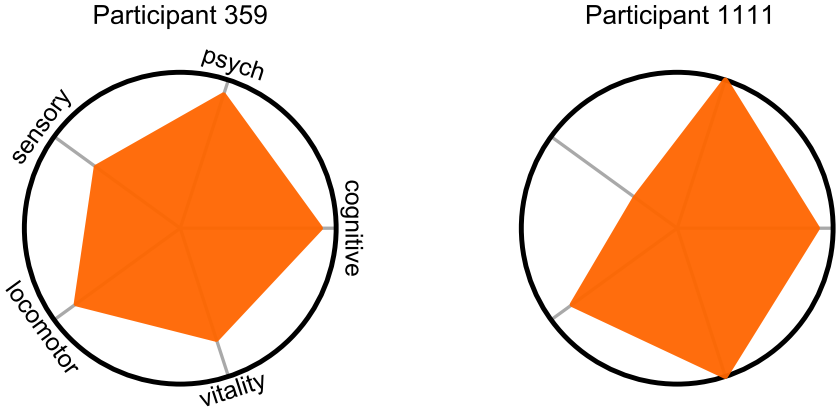
data: $(\sigma_a, \sigma_b) := (\nu \sum_k x_{ak} x_{a,k+1},\ \nu \sum_k x_{bk} x_{b,k+1})$, the test $[\![sgn(\sigma_a - \sigma_b) = sgn(s_a - s_b)]\!]$ is performed, and failures are recorded as violations of the relationship being investigated; (c) The latter step is repeated multiple times, once for each arrangement of the instrument's domains (or vertices order, as presented in Section 4.2.3). These are the results we found:

- Dataset WHOQOL, with $m = 100, d = 4$ (3 arrangements). This setting generates $3\binom{100}{2} = 14,850$ pairs, out of which 1266 were discarded. No violations were found in any of the three arrangements.

- Dataset ELSIO1, with $m = 718, d = 5$ (12 arrangements). This setting generates $12\binom{718}{2} = 3,088,836$ pairs, from which $138,552$ were discarded. The violation rates observed in all arrangements ranged from 0.6% to 0.9%, with weighted average of 0.8%. Example of a pair in which a violation occurs: Participants 359 and 1111 in Figure 35, with $(s_a, s_b) = (4.042, 4.067)$ and $(\sigma_a, \sigma_b) = (1.548, 1.543)$. The pair in this example violates the relationship in 7 out of the 12 arrangements. No pair violates the relationship in all 12 arrangements, but violations occur in all arrangements.

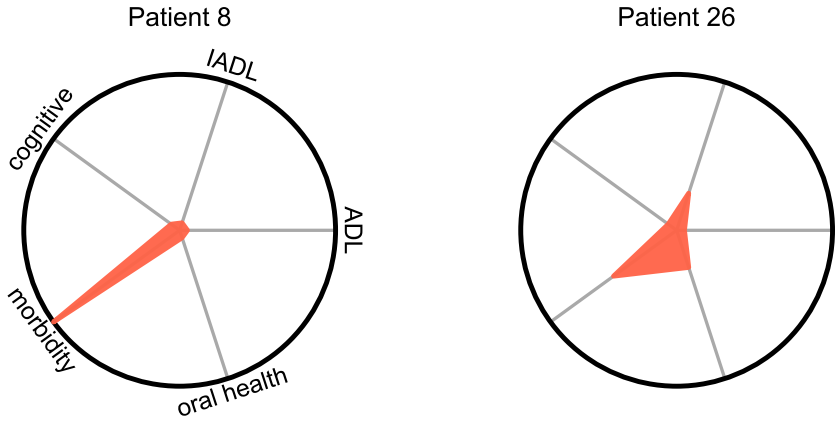
- Dataset AMPIAB, with $m = 510, d = 5$ (12 arrangements). This setting generates $12\binom{510}{2} = 1,557,540$ pairs, out of which $63,708$ were discarded. The violation rates observed in all arrangements ranged from 4.9% to 7.7%, with weighted average of 6.6%. Example of a pair in which a violation occurs: Patients 8 and 26 in Figure 36, with $(s_a, s_b) = (1.217, 1.117)$ and $(\sigma_a, \sigma_b) = (0.059, 0.095)$. Differently from the previous case, the pair in this example violates the relationship in all arrangements. One hypothesis for this larger average violation rate compared to ELSIO1 is that the data fit a congeneric model poorly. As seen in Section 2.4.3, the AMPI/AB instrument achieved low scores in standard unidimensionality tests.

Figure 35 – Two assessments violating the sum-area relationship in elsio1 dataset



Source: The author.

Figure 36 – Two assessments violating the sum-area relationship in ampiab dataset

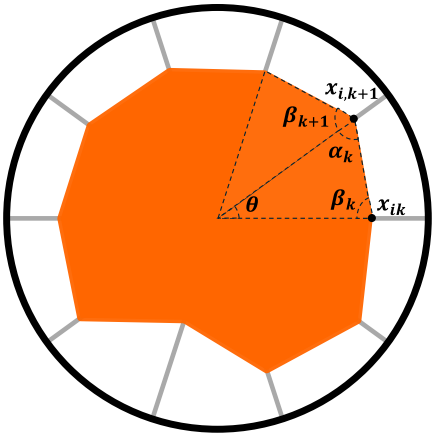


Source: The author.

Based on these results, we conclude that the relationship between sum- and area-scores is much stronger in the WHOQOL and ELSIO1 datasets than in the AMPIAB dataset.

A final word about this exercise in idealisation. Under assumptions a1-a5, the three internal angles of each covering triangle σ_{ik} are determined by the factor loadings. This is to say that, in the diagram in Figure 37, α_k, β_k and θ are determined solely by the factor loadings $\Lambda = (\lambda_0, \ldots, \lambda_{d-1})$. Given that any internal angle of an assessment polygon can be expressed by $(\alpha_k + \beta_{k+1})$, all internal angles are determined similarly. This implies that all assessments in a dataset must induce polygons that are similar — the polygons have the same shape but may vary in size, and the latter is determined solely by η_i. Moreover, Hancock and An (2020) has developed a closed form to estimate Λ directly from the sample covariance matrix of $\mathring{X}$. This result may be useful in the future for the development of clustering techniques for psychometric data based on the Polygrid model.

Figure 37 – Factor loadings determine the internal angles of an assessment polygon



$$\theta = \frac{2\pi}{d}$$

$$\sin\alpha_k = \frac{\lambda_k \sin\theta}{\sqrt{\lambda_k^2 + \lambda_{k+1}^2 - 2\,\lambda_k\,\lambda_{k+1}\cos\theta}}$$

$$\sin\beta_k = \frac{\lambda_{k+1} \sin\theta}{\sqrt{\lambda_k^2 + \lambda_{k+1}^2 - 2\,\lambda_k\,\lambda_{k+1}\cos\theta}}.$$

Source: The author.

The formulas in Figure 37 were developed as follows:

Step 1 - by the law of sines:

$$\frac{\sin\alpha_k}{x_{ik}} = \frac{\sin\theta}{c_{ik}} \implies \sin\alpha_k = \frac{x_{ik}}{c_{ik}}\sin\theta = \frac{\lambda_k\,\eta_i}{c_{ik}}\sin\theta\,,$$
$$\frac{\sin\beta_k}{x_{i,k+1}} = \frac{\sin\theta}{c_{ik}} \implies \sin\beta_k = \frac{x_{i,k+1}}{c_{ik}}\sin\theta = \frac{\lambda_{k+1}\,\eta_i}{c_{ik}}\sin\theta\,.$$

Step 2 - by the law of cosines:

$$\begin{aligned}
c_{ik}^2 &= x_{ik}^2 + x_{i,k+1}^2 - 2\,x_{ik}\,x_{i,k+1}\cos\theta = \\
&= (\lambda_k\,\eta_i)^2 + (\lambda_{k+1}\,\eta_i)^2 - 2\,(\lambda_k\,\eta_i)\,(\lambda_{k+1}\,\eta_i)\cos\theta = \\
&= \lambda_k^2\,\eta_i^2 + \lambda_{k+1}^2\,\eta_i^2 - 2\,\eta_i^2\,\lambda_k\,\lambda_{k+1}\cos\theta = \\
&= \eta_i^2\,(\lambda_k^2 + \lambda_{k+1}^2 - 2\,\lambda_k\,\lambda_{k+1}\cos\theta) \implies \\
c_{ik} &= \eta_i\sqrt{\lambda_k^2 + \lambda_{k+1}^2 - 2\,\lambda_k\,\lambda_{k+1}\cos\theta}\,.
\end{aligned}$$

Step 3 - joining steps 1 and 2:

$$\sin\alpha_k = \frac{\lambda_k\,\eta_i}{c_{ik}}\sin\theta = \frac{\lambda_k}{\sqrt{\lambda_k^2 + \lambda_{k+1}^2 - 2\,\lambda_k\,\lambda_{k+1}\cos\theta}}\sin\theta\,,$$
$$\sin\beta_k = \frac{\lambda_{k+1}\,\eta_i}{c_{ik}}\sin\theta = \frac{\lambda_{k+1}}{\sqrt{\lambda_k^2 + \lambda_{k+1}^2 - 2\,\lambda_k\,\lambda_{k+1}\cos\theta}}\sin\theta\,.$$

Sanity check 1 - letting $\theta = \pi/2$ leads to the Pythagorean setting (as in Figure 34):

$$\sin\alpha_k = \frac{\lambda_k}{\sqrt{\lambda_k^2 + \lambda_{k+1}^2}} = \frac{\lambda_k\eta_i}{\eta_i\sqrt{\lambda_k^2 + \lambda_{k+1}^2}} = \frac{x_{ik}}{\sqrt{x_{ik}^2 + x_{i,k+1}^2}}\,,$$
$$\sin\beta_k = \frac{\lambda_{k+1}}{\sqrt{\lambda_k^2 + \lambda_{k+1}^2}} = \frac{\lambda_{k+1}\eta_i}{\eta_i\sqrt{\lambda_k^2 + \lambda_{k+1}^2}} = \frac{x_{k+1}}{\sqrt{x_k^2 + x_{k+1}^2}}\,.$$

Sanity check 2 - letting $\theta = \pi/2$ and $\lambda_k = 1$ leads to squares:

$$\sin\alpha_k = \frac{\lambda_k}{\sqrt{\lambda_k^2 + \lambda_{k+1}^2}} = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2} \implies \alpha_k = \frac{\pi}{4}\,,$$
$$\sin\beta_k = \frac{\lambda_{k+1}}{\sqrt{\lambda_k^2 + \lambda_{k+1}^2}} = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2} \implies \beta_k = \frac{\pi}{4}\,.$$

Thus, it must be the case that $(\alpha_k + \beta_{k+1}) = \dfrac{\pi}{2}$, $k = 0 \ldots d-1$.