Novel strategies for identifying and addressing mental health and learning disorders in school-age children

Kseniia Konishcheva

Université Paris Cité
ED Frontières de l'Innovation en Recherche et Éducation (474)
Inserm SEED Évolution et ingénierie de systèmes dynamiques (U 1284, UMR-S 1284)

Novel Strategies for Identifying and Addressing Mental Health and Learning Disorders in School-Age Children

Kseniia KONISHCHEVA

Une thèse de doctorat en Neuroscience et Troubles Neuronaux
Dirigée par Dr. Ariel B. Lindner et Dr. Arno Klein

Présentée et soutenue publiquement le 20/12/2023 devant un jury composé de :

Dr. Ariel B. Lindner, Directeur de recherche, Université Paris Cité/INSERM, *Directeur*
Dr. Arno Klein, Director, Assistant Professor, Child Mind Institute, *Directeur*
Pr. Catherine Lord, Full Professor, University of California, *Présidente du jury*
Dr. Satrajit Ghosh, Principal Research Scientist, Assistant Professor, Massachusetts Institute of Technology, *Rapporteur & Examinateur*
Pr. Yasser Khazaal, Full Professor, Lausanne University, *Rapporteur & Examinateur*
Dr. Nicole Landi, Associate Professor, University of Connecticut, *Examinatrice*
Dr. Amy Margolis, Associate Professor, Columbia University, *Invitée*
Dr. Bennett Leventhal, Professor Emeritus, University of California San Francisco, *Invité*

Resumé

Titre: Nouvelles stratégies d'identification et de gestion des troubles de la santé mentale et des troubles de l'apprentissage chez les enfants d'âge scolaire.

Mots clés: Psychiatrie de l'enfant, Dépistage universel de la santé mentale, Santé mentale en milieu scolaire, Évaluation de la santé mentale, Apprentissage automatique (Machine Learning), Sélection des caractéristiques

La prévalence des troubles de la santé mentale et de l'apprentissage chez les enfants d'âge scolaire est une préoccupation croissante. Pourtant, il existe un délai important entre l'apparition des symptômes et l'orientation vers une intervention, ce qui contribue à des problèmes sur le long terme pour les enfants concernés. Le système actuel de santé mentale est fragmenté : les enseignants ont une connaissance précieuse du bien-être de leurs élèves, mais peu de connaissances en matière de santé mentale, tandis que les cliniciens ne rencontrent souvent que les cas les plus graves.

La mise en œuvre incohérente des programmes de dépistage dans les écoles, principalement en raison de contraintes de ressources, suggère la nécessité de solutions plus efficaces. Cette thèse présente deux nouvelles approches visant à améliorer la santé mentale et les résultats d'apprentissage des enfants et des adolescents.

La première approche utilise une méthode "data-driven", en tirant parti de l'ensemble de données du Healthy Brain Network, qui contient plus de 50 évaluations et leurs réponses, des diagnostics, et des scores de tâches cognitives de milliers d'enfants. À l'aide de techniques de machine learning, des sous-ensembles d'éléments ont été identifiés pour prédire les diagnostics courants de santé mentale et de troubles de l'apprentissage. L'approche a démontré des performances prometteuses, offrant une potentielle utilité pour la détection des troubles mentaux et des troubles de l'apprentissage. En outre, notre approche constitue un point de départ utile pour les chercheurs qui souhaitent appliquer notre méthode sur de nouveaux ensembles de données.

La deuxième approche est un framework visant à améliorer la santé mentale et les résultats d'apprentissage des enfants en relevant les défis auxquels sont confrontés les enseignants dans les classes hétérogènes. Ce framework permet aux enseignants de créer des stratégies d'enseignement sur mesure basées sur les besoins de chaque élève et, si nécessaire, de

suggérer une orientation vers des soins cliniques. La première étape du framework est un outil conçu pour évaluer le bien-être et le profil d'apprentissage de chaque élève. FACETS est un questionnaire de 60 points élaboré dans le cadre de partenariats avec des enseignants et des cliniciens. L'acceptation des enseignants et les propriétés psychométriques de FACETS sont étudiées. Une étude pilote préliminaire a démontré l'acceptation générale de FACETS par les enseignants.

En conclusion, cette thèse présente un cadre permettant de combler les lacunes en matière de détection et de soutien des troubles mentaux et des troubles de l'apprentissage chez les enfants d'âge scolaire. De futures études permettront de valider et d'affiner nos outils, offrant ainsi des interventions plus opportunes et plus efficaces pour améliorer le bien-être et les résultats d'apprentissage des enfants dans divers contextes éducatifs.

Abstract

Title: Novel Strategies for Identifying and Addressing Mental Health and Learning Disorders in School-Age Children

Keywords: Child Psychiatry, Universal Mental Health Screening, School-based Mental Health Interventions, Mental Health Assessment, Machine Learning, Feature Selection

The prevalence of mental health and learning disorders in school-age children is a growing concern. Yet, a significant delay exists between the onset of symptoms and referral for intervention, contributing to long-term challenges for affected children. The current mental health system is fragmented, with teachers possessing valuable insights into their students' well-being but limited knowledge of mental health, while clinicians often only encounter more severe cases.

Inconsistent implementation of existing screening programs in schools, mainly due to resource constraints, suggests the need for more effective solutions. This thesis presents two novel approaches for improvement of mental health and learning outcomes of children and adolescents.

The first approach uses data-driven methods, leveraging the Healthy Brain Network dataset which contains item-level responses from over 50 assessments, consensus diagnoses, and cognitive task scores from thousands of children. Using machine learning techniques, item subsets were identified to predict common mental health and learning disability diagnoses. The approach demonstrated promising performance, offering potential utility for both mental health and learning disability detection. Furthermore, my approach provides an easy-to-use starting point for researchers to apply the method to new datasets.

The second approach is a framework aimed at improving the mental health and learning outcomes of children by addressing the challenges faced by teachers in heterogeneous classrooms. This framework enables teachers to create tailored teaching strategies based on identified needs of individual students, and when necessary, suggest referral to clinical care. The first step of the framework is an instrument designed to assess each student's well-being and learning profile. FACETS is a 60-item scale built through partnerships with teachers and clinicians. Teacher acceptance and psychometric properties of FACETS are investigated. Preliminary pilot study demonstrated overall acceptance of FACETS among teachers.

In conclusion, this thesis presents a framework to bridge the gap in detection and support of mental health and learning disorders in school-age children. Future studies will further validate and refine the tools, offering more timely and effective interventions to improve the well-being and learning outcomes of children in diverse educational settings.

Résumé substantial

La prévalence des troubles mentaux et des troubles de l'apprentissage chez les enfants en âge scolaire est de plus en plus préoccupante. Des recherches menées dans 27 pays, axées sur des études reposant sur des méthodologies solides, suggèrent qu'environ 13 % des enfants et des adolescents sont affectés par un trouble mental à un moment donné. Une enquête de 2004 indique que la moitié des troubles mentaux apparaissent avant l'âge de 14 ans, et 75 % avant l'âge de 24 ans. Les résultats de méta-analyses récentes confirment l'hypothèse que la plupart des troubles mentaux apparaissent pendant l'adolescence.

Les problèmes de santé mentale sont liés à diverses conséquences négatives, notamment l'échec scolaire et professionnel, la sous-performance, les difficultés financières, l'altération des relations sociales, l'accès limité aux soins de santé et une réduction potentielle de l'espérance de vie pouvant aller jusqu'à vingt ans.

Les enfants souffrant d'une maladie mentale encourent un risque accru d'échec scolaire et de décrochage. Les difficultés scolaires, à leur tour, sont en corrélation avec d'autres conséquences négatives dans la vie de tous les jours, telles qu'un revenu plus faible et une santé plus fragile.

Si les problèmes de santé mentale chez les jeunes ne sont pas traités, ils persistent souvent à l'âge adulte, entraînant les conséquences négatives susmentionnées. Les antécédents de troubles mentaux pendant l'enfance et l'adolescence s'avèrent être un prédicteur plus puissant des conséquences néfastes dans le quotidien, que la présence actuelle d'un trouble mental.

Malgré les répercussions importantes que les problèmes de santé mentale non traités peuvent avoir sur les personnes concernées, beaucoup d'entre elles ne reçoivent pas les soins appropriés,le traitement subissant souvent un retard important. L'écart entre les premiers signes d'un trouble et le début du traitement peut s'étendre sur plusieurs années et, dans certains cas, les personnes ne reçoivent aucun traitement. Cette disparité est observée même dans les pays développés, où les troubles mentaux sont souvent sous-traités par rapport aux affections physiques, bien qu'ils entraînent des niveaux d'invalidité similaires. En outre, une grande partie de la population qui signale des problèmes de santé mentale bénéficie de services inadéquats, une faible proportion d'entre eux cherchant de l'aide par rapport à la prévalence estimée dans l'ensemble de la population. Cette situation est

exacerbée par un âge d'apparition plus précoce, ce qui entraîne des retards supplémentaires dans le traitement. Les pays en développement sont confrontés à une situation encore plus difficile.

Lorsque les personnes concernées recherchent une aide professionnelle, la majorité d'entre elles la trouvent bénéfique. Dans la plupart des cas, l'identification d'un trouble dépend de la recherche d'aide, soit de la part de la personne affectée elle-même, soit de la part d'une tierce personne telle que des amis, des enseignants, des parents ou des services d'urgence. Le manque généralisé de sensibilisation à la santé mentale contribue à la difficulté de reconnaître les symptômes, à la fois chez soi et chez les autres. Ce problème empêche une intervention rapide et est particulièrement évident chez les enfants, dont les problèmes de santé mentale passent souvent inaperçus ou sont considérés comme normaux pour leur âge. Même lorsque des symptômes sont identifiés, de nombreuses personnes hésitent à demander de l'aide, attendant souvent des années avant de le faire.

L'identification et la prise en charge précoce des symptômes peuvent réduire le risque d'évolution vers des troubles à part entière, contribuant ainsi au bien-être général de la communauté et réduisant le fardeau économique de la maladie mentale. Une intervention précoce peut permettre d'éviter une invalidité à vie dans le cas de certains troubles.

Les écoles offrent un environnement unique pour la détection des symptômes précoces des troubles mentaux chez les élèves. Les enseignants, qui observent les élèves dans un contexte différent de celui des parents et des cliniciens de soins primaires, sont souvent bien placés pour identifier les signes de maladie mentale. Alors que les écoles constituent déjà une source majeure de soutien en matière de santé mentale, en particulier pour les jeunes des minorités rurales, l'approche actuelle qui consiste à "attendre l'échec" retarde les évaluations et les services spécialisés jusqu'à ce que les élèves soient confrontés à des difficultés scolaires ou sociales prolongées.

Une autre approche consiste en un dépistage universel de la santé mentale dans les écoles, qui s'est avéré faisable et efficace pour identifier les troubles, en particulier parmi les groupes traditionnellement sous-diagnostiqués. Malheureusement, malgré les encouragements, de nombreuses écoles ne mettent pas systématiquement en oeuvre le dépistage universel de la santé mentale en raison de divers obstacles, notamment un manque de sensibilisation, un accès limité aux instruments de dépistage, des ressources financières insuffisantes et un manque de connaissances sur la manière d'aider les élèves ayant des besoins identifiés.

Les obstacles courants à la mise en œuvre d'un dépistage universel de la santé mentale sont notamment la méconnaissance des programmes existants, l'accès limité aux instruments de dépistage et l'insuffisance des ressources financières. Les enseignants et les administrateurs scolaires peuvent manquer de connaissances sur la manière d'aider les élèves ayant des besoins identifiés et ne pas connaître les voies d'orientation. En outre, il n'existe pas de norme universellement acceptée pour le dépistage universel de la santé mentale, et les instructions pratiques pour les écoles sont limitées.

La mise en œuvre irrégulière des programmes de dépistage existants dans les écoles, principalement en raison de contraintes de ressources, souligne la nécessité de trouver des solutions plus efficaces. Cette thèse présente deux approches innovantes visant à améliorer la santé mentale et les résultats d'apprentissage des enfants et des adolescents.

La première approche fait appel à des méthodes basées sur les données, en utilisant l'ensemble de données du Healthy Brain Network (HBN), qui comprend des réponses au niveau des éléments de plus de 50 évaluations, des diagnostics et des scores de tâches cognitives de milliers d'enfants. Grâce à des techniques d'apprentissage automatique, des sous-ensembles spécifiques d'éléments ont été identifiés pour prédire les diagnostics courants de santé mentale et de troubles de l'apprentissage.

Nous avons démontré que nos modèles d'apprentissage automatique étaient plus performants que n'importe lesquels des évaluations HBN existantes. L'inclusion d'éléments provenant uniquement d'évaluations non propriétaires ou de rapports de parents n'a pas eu d'impact significatif sur les performances, ce qui pourrait s'expliquer par les similitudes entre les questions dans la vaste liste. Cette constatation est étayée par l'observation selon laquelle l'intégration d'un plus grand nombre d'évaluations n'a pas entraîné d'amélioration substantielle des performances.

Notre approche a montré des performances prometteuses en utilisant l'ensemble de données HBN, ce qui en fait une ressource précieuse pour les futurs chercheurs cherchant à développer des instruments de dépistage plus efficaces.

Après avoir obtenu le consentement des auteurs de l'évaluation originale, les chercheurs peuvent utiliser les sous-ensembles d'items proposés pour créer et valider de nouvelles évaluations. Celles-ci peuvent être évaluées soit par une simple notation sommaire, soit en

utilisant les modèles d'apprentissage automatique entraînés, à condition que la population évaluée soit comparable à la population de l'ensemble de données. Les valeurs de

performance des modèles d'apprentissage automatique offrent des estimations préliminaires de l'efficacité du nouveau filtre dans une population similaire à celle de l'ensemble de données.

Bien que nos sous-ensembles aient montré de bonnes performances dans l'ensemble de données HBN, la validation des sous-ensembles d'éléments identifiés dans les populations cibles est essentielle.

Nous offrons aux chercheurs un point de départ accessible sur GitHub, où ils peuvent appliquer notre méthode à n'importe quel ensemble de données afin d'identifier les éléments de dépistage des troubles qui les intéressent.

Pour l'avenir, nous suggérons d'explorer la sélection de caractéristiques multi-labels afin d'identifier des sous-ensembles d'items pour plusieurs troubles simultanément. La stratification de l'ensemble d'apprentissage par âge et le développement de modèles distincts pour les différentes tranches d'âge pourraient améliorer la précision du dépistage dans les différents groupes d'âge.

Pour faciliter l'application de nos modèles en tant que méthode de notation, nous recommandons de développer une interface conviviale pour l'administration et la notation de l’évaluation..

La deuxième approche implique un cadre conçu pour améliorer la santé mentale et les résultats d'apprentissage des enfants en relevant les défis dans les classes hétérogènes, en fournissant aux écoles les ressources nécessaires pour évaluer systématiquement les besoins des élèves et y répondre. Dans un premier temps, la collaboration avec les enseignants, les chercheurs et les cliniciens a conduit à l'élaboration d'un outil d'évaluation. Cet outil offre une vue d'ensemble du fonctionnement d'un enfant, englobant la santé mentale, les capacités d'apprentissage, le comportement, la cognition et les émotions. Il établit un langage commun pour la communication et la collaboration entre les éducateurs, les parents, les experts cliniques et les autres parties prenantes. Le questionnaire est destiné aux enseignants qui cherchent à mieux comprendre les profils des élèves (points forts et besoins) afin de leur offrir un soutien adapté. FACETS vise à dépister les problèmes de santé

mentale, d'apprentissage, de comportement et autres, ce qui permet d'intervenir en temps utile et d'orienter les élèves vers des services spécialisés. Les enseignants sont encouragés à évaluer tous les élèves de leur classe afin d'éviter les préjugés et d'identifier ceux qui ont besoin d'aide et dont les symptômes sont moins évidents ou perturbateurs. La faisabilité de FACETS a été assurée, en tenant compte de la capacité des enseignants à comprendre et à observer tous les comportements évalués, tout en minimisant le temps nécessaire. La plupart des éléments de FACETS utilisent une échelle visuelle analogique, les deux extrémités représentant les faiblesses et le milieu indiquant un comportement typique pour le niveau de développement attendu de l'enfant. Les questions sont dérivées des principaux symptômes des troubles mentaux et des domaines de préoccupation identifiés par les éducateurs. Ils ne sont pas spécifiques à un seul trouble, mais apparaissent couramment dans les salles de classe comme des obstacles familiers à l'apprentissage. Au cours de l'élaboration de FACETS, les enseignants ont hiérarchisé les éléments en fonction de leur prévalence et de leur impact sur l'apprentissage. Le questionnaire est divisé en dix sections évaluant différents aspects du bien-être et de l'apprentissage, la dernière section contenant des questions dichotomiques indiquant les problèmes nécessitant une attention particulière.

L'objectif de la recherche actuelle est d'établir à la fois l'acceptation par les enseignants et les propriétés psychométriques de FACETS.

Une étude pilote a été entreprise dans une école pour évaluer l'acceptation de FACETS par les enseignants et recueillir des données psychométriques préliminaires, y compris la fiabilité inter-juges et test-retest. L'acceptation par les enseignants a été déterminée par le biais de discussions de groupe et d'une enquête anonyme.

Les résultats des sessions des groupes de discussion ont montré que FACETS était globalement accepté et ont identifié des domaines à améliorer.

La fiabilité inter-évaluateurs et test-retest s'est avérée faible à moyenne, avec une taille d'échantillon insuffisante pour obtenir des intervalles de confiance significatifs pour la plupart des items. Aucune modification significative de la fiabilité n'a été observée entre les deux parties de l'enquête FACETS, ce qui suggère l'absence de lassitude de la part des évaluateurs. La fiabilité test-retest a diminué avec le temps écoulé entre les administrations, soulignant la nécessité d'un interval de temps fixe dans les études futures.Un regroupement hiérarchique des données relatives aux enseignants a montré des regroupements d'items attendus, suggérant que les enseignants comprenaient bien les items.

L'étude pilote a donné lieu à des révisions de l'instrument, notamment l'élaboration d'une version française, la reformulation des questions et des options de réponse sur la base des commentaires des enseignants, ainsi qu'une nouvelle plateforme en ligne.

Une étude plus large en cours examine la fiabilité de la version révisée de FACETS. La première phase, qui vise à établir l'acceptation par les enseignants, est terminée, et la deuxième phase est prévue pour novembre 2023.

Parallèlement, une étude de validité est en cours, évaluant les éléments de FACETS par rapport aux évaluations de routine dans un service de pédopsychiatrie. La phase d'acceptation par les cliniciens est terminée et des ajustements mineurs ont été apportés en fonction des réactions. La collecte de données pour l'évaluation de la validité est en cours.

FACETS fait désormais partie du cadre LISA, qui a reçu une subvention de 2 millions d'euros sur cinq ans, afin de s'étendre à plus de 230 établissements pilotes et de s'intégrer dans le système éducatif national pour des études plus vastes et diversifiées visant à confirmer les propriétés psychométriques de FACETS.

En résumé, cette thèse vise à remédier à la disparité existante dans l'identification et l'assistance des troubles de la santé mentale et de l'apprentissage chez les enfants d'âge scolaire. Les efforts de recherche ultérieurs se concentreront sur la validation et l'amélioration des outils développés, dans le but de fournir des interventions plus rapides et plus efficaces qui améliorent le bien-être général et les résultats scolaires des enfants dans une variété d'environnements.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Ariel Lindner and Arno Klein, for their support, guidance, and encouragement throughout my Ph.D. journey. Their insights and mentorship have been invaluable.

A special thanks to my mentors and colleagues: Michael Milham, Bennett Leventhal, Maki Koyama, Anirudh Krishnakumar, Elie Rotenberg, Richard Delorme, Joel Swendsen, Mauricio Scopel Hoffmann, Pernille Brams, and Laure Mourgue d'Algue.

A heartfelt thank you goes to the iféa school network, its founder Elie Rotenberg, and the dedicated staff and teachers who played a vital role in the conception and development of the FACETS and the LISA framework.

I extend my appreciation to the members of the jury, Satrajit Ghosh, Yasser Khazaal, Catherine Lord, Nicole Landi, Amy Margolis, and Bennett Leventhal, for their time and willingness to provide thoughtful evaluation of my work.

I am grateful to those who provided valuable feedback and assistance, including Satrajit Ghosh and his team, Joshua Vogelstein, Sambit Panda, Nicole Landi, Elena Grigorenko, Giovanni Salum, and Rebecca Neuhaus.

A sincere thank you to the Child Mind Institute staff for their warm welcome, feedback, and assistance. I am also thankful to the Learning Planet Institute staff, particularly Lionel Deveaux, Beatrice Herbin, Dule Misevic, Camille Gaulon, and Elodie Kaslikowski, for their support in facilitating my work.

I appreciate the camaraderie and assistance from fellow Ph.D. students who shared insights and provided a supportive environment.

Lastly, heartfelt thanks to my parents, my husband, and my friends, for their encouragement and support throughout this journey.

Table of contents

List of Figures

List of Tables

Preface

This thesis addresses the critical issue of mental health and learning disorders among school-age children. The prevalence of such disorders has become a growing concern, and timely detection is crucial for effective support. In this preface, I provide an overview of the structure, and objectives of this thesis.

My research is centered on addressing the challenges faced by educators, mental health professionals, and, most importantly, the children affected by these disorders, using theory-driven and data-driven approaches. The overarching goal is to enhance the detection and support, ultimately leading to an improvement of the overall well-being of children.

Chapter 1 introduces categorical and dimensional approaches to mental health, discussing the prevalence of these disorders, and the negative outcomes associated with delayed intervention. This chapter also identifies challenges addressed in my research by highlighting obstacles to help-seeking and receiving treatment, introducing the potential of systematic mental health screening. Additionally, it described the properties of psychological tests and the development process of screening tools, detailing methodological approaches relevant to the subsequent chapters.

In Chapter 2, titled "Optimizing Item Selection for Psychiatric and Learning Disorder Screening Using Machine Learning," I leverage the Healthy Brain Network dataset and machine learning techniques to define and test a novel, data-driven approach for creating new mental health assessments.

In Chapter 3 introduces the "Learning, Integration, Support, and Awareness (LISA) Framework" which is designed for systematic early identification and management of mental health and learning difficulties in schools. In this chapter I examine the psychometric properties of FACETS, a new 60-item instrument designed to assess each student's well-being and learning profile.

Each chapter's introduction provides relevant background information, as well as challenges and research aims. Then, methods and results are presented, followed by discussions of ongoing work and future directions.

Finally, in Chapter 4, I summarize the outcomes of each chapter, and discuss how the integration of the two approaches contributes to improvement of identification and management of mental health and learning disorder in children. Additionally, I contextualize my research results within the framework of the recent work by other teams and organizations, and discuss the potential overlap between the two approaches for future research.

1. Introduction

1.1. Categorical approaches to mental health

Psychological functioning is notoriously difficult to measure (Vessonen, 2020). Unlike physical diagnoses, most psychological constructs cannot be measured directly (Fried & Flake, 2018) and rely on patient reports and observations (Balogh et al., 2015; Flake & Fried, 2019; Fried & Flake, 2018). The word construct in this context refers to the concept that needs to be measured.

There is more disagreement on the nature or mental disorder diagnosis than on diagnosis in other branches of medicine (Kessler, 2007). It has been noted that mental disorders are fundamentally different from other clinical entities. Psychiatric symptoms are not indicators of disorders, but themselves constitute a disorder – i.e. you cannot have asymptomatic mental disorder (Roefs et al., 2022). Mental health diagnosis relies on consensus-based diagnostic manuals, due to the lack of biomarkers and known etiology for most disorders (Yan et al., 2022).

The two most commonly used systems for codifying and standardizing psychiatric diagnoses are the Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 2013) and the International Classification of Diseases (ICD; World Health Organization, 2019). The ICD started as a classification system of causes of death used for mortality statistics. Currently it is used as an international manual classifying disease for both clinical and statistical purposes. In its sixth revision published in 1949 by the World Health Organization (WHO) a new chapter was included dedicated to psychiatric disorders (Hirsch et al., 2016).

The first version of DSM was published by APA (American Psychiatric Association, American Medico-Psychological Association at the time) in 1952, three years after the publication of ICD-6, and was in part inspired by it. ICD-8 and DSM-II were developed in close collaboration between WHO and APA, resulting in a nearly identical classification. The third version of DSM (DSM-III) released in 1980 included explicit diagnostic criteria based on observable signs and patient-reported symptoms (e.g. 3 out of 5 symptoms need to be present for a diagnosis), which made diagnosis given by different clinicians more reliable (Clark et al., 2017; KATSCHNIG, 2010; Parnas, 2015). Currently ICD is commonly used in healthcare in Europe,

while DSM is more commonly used by clinicians in the United States, and in research worldwide (KATSCHNIG, 2010).

1.2. Dimensional approaches to mental health

The dimensional nature of many mental health symptoms has been explored since the 1920s (Angst, 2007), and has since then been corroborated by recent genetic and imaging studies (J. S. Anderson et al., 2019; Hägele et al., 2015).

As opposed to the categorical approach to the diagnosis based on a list of diagnostic criteria, the dimensional approach can be compared to hypertension, where a pathological and non-pathological range exists within a continuous measure, and treatment is aimed at reducing the severity of symptoms within this continuum (Lubke et al., 2009).

The latest versions of the DSM and ICD integrated the dimensional approach for some disorders (Gaebel et al., 2020; Regier et al., 2013). The overall categorical classification was nonetheless conserved due to insufficient scientific consensus on which dimensions should be used (Mitropoulos, 2018).

1.3. Prevalence of mental disorders

The introduction of explicit diagnostic criteria in DSM-III in 1980s provided an opportunity to reliably estimate the prevalence of mental disorders in the general population (Polanczyk et al., 2015).

Prevalence of a disorder refers to an estimate of the proportion of people affected by the disorder in a particular population in a particular time period. *Point prevalence* is the prevalence at a particular point in time, *period prevalence* is the prevalence at any point during a particular time period (e.g., past year). *Lifetime prevalence* refers to the proportion of individuals who were affected at some point in their life (*What Is Prevalence?*, n.d.). Two common sources of prevalence estimates are administrative data and cross-sectional studies.

Administrative data includes data routinely collected by government agencies and insurance companies. Administrative data is attractive for epidemiology research because of the large volumes of data collected systematically over years for a large portion of the population

(Ward, 2013). However, administrative data is not ideal to measure prevalence of mental disorders, as it only includes people who are in contact with mental health services (Duncan et al., 2022). As will be discussed further, people tend to seek help for mental health symptoms less often than for physical problems (Kessler, 2007). Another problem with using administrative data for estimating mental disorder prevalence is the lack of standardized approach for establishing whether an individual is affected by the disorder (Duncan et al., 2022). Clinicians have different levels of experience and approaches to diagnosis (Dattani, 2023). In some cases, clinicians intentionally record a "wrong" diagnosis to circumvent insurance policy, for example to obtain reimbursement for treatment that would not be eligible for reimbursement otherwise (Aboraya, 2007). Administrative data is generally not reliable when comparing prevalence between countries, due to differences in administrative data collection practices (Dattani, 2023).

Cross-sectional studies are an alternative approach for prevalence estimation. A cross-sectional study is a study on a representative sample from a population at a single point in time (Setia, 2016, p. 3) (as opposed to a longitudinal study, where data is collected at several points in time for the same sample). To estimate prevalence of a disease, researchers randomly select a representative sample from the population, then the participants are sorted into two groups, depending on whether they have a disorder in question, and the number of affected individuals is counted and divided by the total number of people in the sample (Ward, 2013). For reliable prevalence estimates the sample should be sufficiently large and diverse. Estimating prevalence of more rare disorders require even larger sample sizes (Ward, 2013). Cross-sectional studies address the limitations of studies based on administrative data by including individuals who do not seek treatment from mental health services, and by using standardized assessment protocols (Dattani, 2023).

Despite the advantages of the cross-sectional studies for prevalence estimation, the prevalence numbers for the same disorders vary between different studies. The sources of these variations include the methodology for selecting the study sample, and different approaches to establishing the diagnoses (Brauner & Stephens, 2006; Ferrari et al., 2013). Even for physical disorders it is often difficult to establish a definitive diagnosis (Enøe et al., 2000). Epidemiological studies often rely on tests that differ from what a clinician would use to establish a diagnosis (Silva, 1999). For many disorders there is no "gold standard" diagnostic test that can be used to definitively establish diagnosis (Lewis & Torgerson, 2012). This problem is even more pronounced for psychiatric disorders, where there is more disagreement about the criteria, or even about the existence of some diagnoses (Kessler, 2007).

The first large-scale study attempting to estimate nationwide prevalence of mental disorders was the Epidemiologic Catchment Area (ECA) Study, conducted in the 1980s in the US. The researchers used a fully-structured diagnostic interview that was conducted by non-clinicians. *Clinical reappraisal interviews* were conducted to confirm that the interviews produced similar results to independent diagnostic interviews conducted by clinicians, showing favorable results in clinical settings, but significantly lower results in community settings (Leeman, 1999). In the late 1990s, the methodology established by the ECA study was used by WHO for the World Mental Health Survey Initiative (WMH) to estimate the prevalence of mental disorders in multiple countries. So far the survey has been conducted in 30 countries (*The World Mental Health Survey Initiative*, n.d.).

A report including results from 17 WMH showed estimated lifetime prevalence of any mental disorder to vary between countries, the highest being 47.4% in the United States, and the lowest of 12% in Nigeria (Figure 1). Between half the population in some countries and one-fifth of the population in others were estimated to be at risk of having a mental disorder at some point during their life (KESSLER et al., 2007). The WMH surveys have also estimated the age of onset of mental disorders. Unlike the discrepancies in prevalence rates, the age of onset was relatively stable across countries (KESSLER et al., 2007).

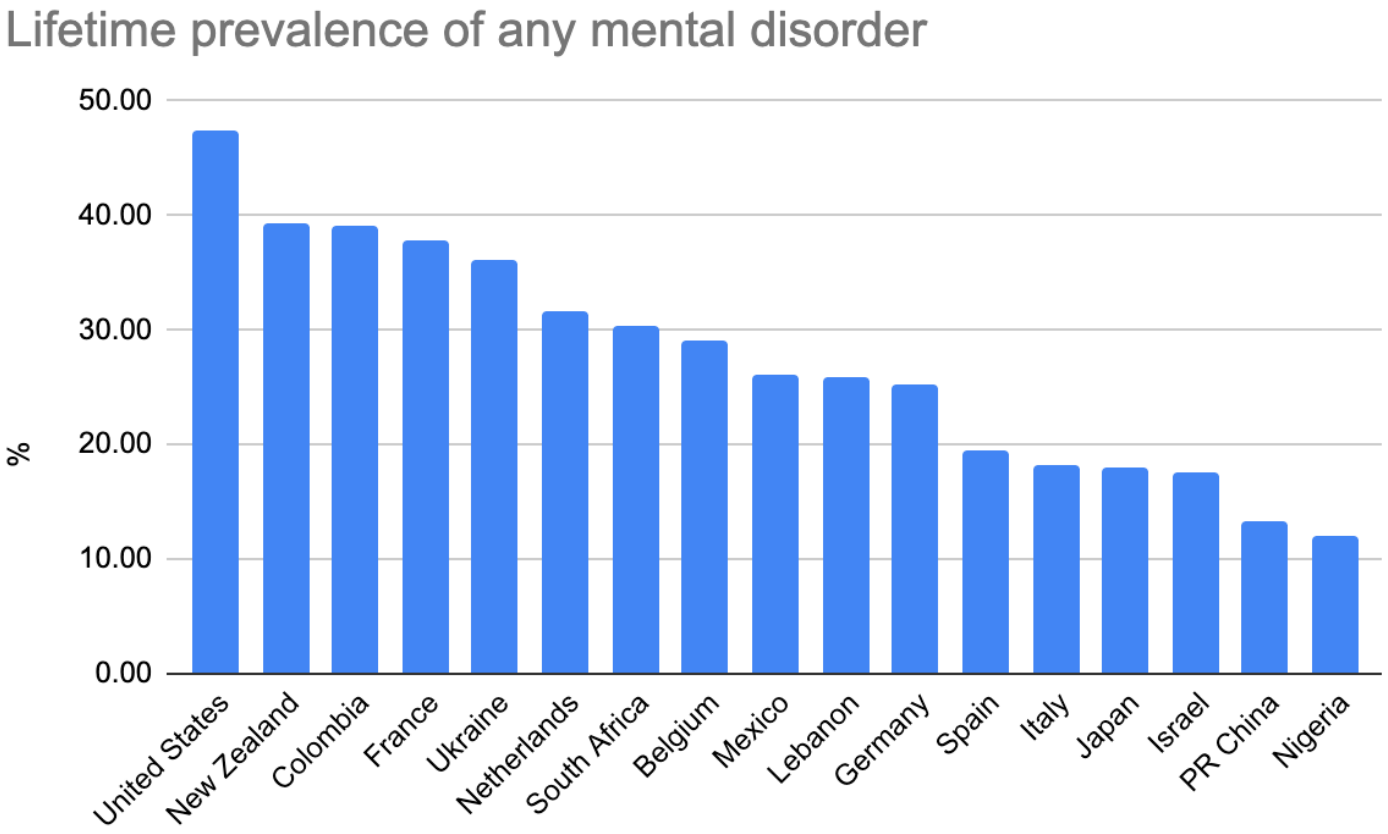

Figure 1: Lifetime prevalence of any mental disorder according to the World Mental Health survey (KESSLER et al., 2007).

There is no equivalent of the WMH study for estimating mental disorder prevalence in children and adolescents across multiple countries using consistent methodology with a standardized diagnostic measure. A systematic review of community studies from 27 countries that only included studies with sound methodologies estimated that around 13% of children and adolescents are affected by a mental disorder at any given moment (Polanczyk et al., 2015). A 2004 survey estimated that half of mental disorders start before the age of 14, and 75% before the age of 24 (Kessler et al., 2005). A recent meta-analysis confirms that most mental disorders start during adolescence (Solmi et al., 2022).

One of the findings of the ECA study was the fact that about half the people with one disorder also have another disorder (Kessler, 2011). The presence of two or more disorders in a single individual is referred to as *comorbidity*. European studies report high comorbidity rates as well. Individuals with comorbid disorders show higher disability and higher rates of help seeking. The high comorbidity rates inspired discussions regarding validity of the boundaries between disorders in the current diagnostic system (Merikangas & Kalaydjian, 2007).

1.4. Negative outcomes of mental disorders

Mental health problems are associated with numerous negative outcomes including academic and job failure and underachievement (Bruns et al., 2016; de Graaf et al., 2008; Kessler, 2007; Kooij, 2012; O'Connor et al., 2018; WANG et al., 2007), financial difficulties (Ormel et al., 2017), impacting social relationships (Bruns et al., 2016; WANG et al., 2007), access to healthcare (N. H. Liu et al., 2017), and decreasing life expectancy for up to twenty years (Fusar-Poli et al., 2021).

Children affected by mental illness are at risk of academic underachievement and school dropout (Bruns et al., 2016; Dalsgaard et al., 2020; O'Connor et al., 2018). Academic underachievement, in turn, is associated with multiple negative life outcomes such as lower income, and worse health outcomes (Dalsgaard et al., 2020).

If mental health problems in youth are not promptly addressed, they often persist into adulthood, resulting in aforementioned negative outcomes (Bruns et al., 2016; Fusar-Poli et al., 2021). The majority of people with a current disorder received a first diagnosis before the age of 18 (De Girolamo et al., 2012). History of mental disorders during childhood and adolescence is a stronger predictor for multiple negative life outcomes than currently having a mental disorder, with some outcomes predicted only by past disorder when using both current and past disorder as predictors (Ormel et al., 2017).

Despite the serious consequences unaddressed mental health issues pose for the affected individuals, many do not receive appropriate care, or receive it after a long delay. It can take years between the first signs of a disorder and initiation of treatment (Dagani et al., 2017), if the treatment is received at all (Ginsberg et al., 2014; Merikangas et al., 2013). In many developed countries, mental disorders are undertreated compared to physical disorders, even when leading to a similar level of disability (Merikangas et al., 2013; Ormel et al., 2008). The proportion of the population that report receiving mental health services is low compared to the estimated prevalence in the general population (Colizzi et al., 2020; Hagell et al., 2021; Merikangas et al., 2009; Moro & Brison, 2016). Earlier age of onset increases delay in treatment even further. (De Girolamo et al., 2012). The situation is even worse in developing countries (Liao et al., 2023). When affected individuals do receive professional help, most find it helpful (Mandalia et al., 2018).

1.5. Obstacles to seeking help

In most cases, identification of a disorder relies on help-seeking, either from the affected individual themself or from a third-party (friends, teachers, parents, emergency services) (MacDonald 2018, MacDonald 2020, Yamasaki 2016, Johnston 2019, Salaheddin 2016).

Mental health awareness is low among most populations (McGorry 2014), which prevents people from recognizing symptoms in themselves or other people (MacDonald et al., 2021; McGorry et al., 2014a; Milgrom & Gemmill, 2014; Schomerus et al., 2019; Tunks et al., 2023; Yamasaki et al., 2016). Most mental health problems in children remain unrecognized by their parents, or dismissed as being normal for the child's age (Johnston & Burke, 2020).

Even when symptoms are recognized, people often do not seek help, or do so only years later (McGorry et al., 2014b; Savage et al., 2016). Help-seeking rates are low even in countries with developed mental health services (Schomerus et al., 2019). Several reasons for this have been reported, including cost, transportation issues (especially in rural areas), privacy concerns, and reluctance to talk about emotional concerns (MacDonald et al., 2021; McGorry et al., 2014a; Salaheddin & Mason, 2016). Sometimes the symptoms in questions themselves are the reason for delayed help-seeking, such as low energy levels associated with depression (Milgrom & Gemmill, 2014). Previous negative experiences with mental health services delay help-seeking in the future (MacDonald et al., 2021).

A big barrier to help-seeking is stigma and embarrassment associated with mental health problems, both among affected individuals and their caregivers (MacDonald et al., 2021; Salaheddin & Mason, 2016; Yamasaki et al., 2016). Both perceived stigma (fear of being judged by others (Tesfaw et al., 2020)), and self-stigma (internalizing the judgmental attitude prevalent in society and experiencing shame about your condition (*Stigma, Prejudice and Discrimination Against People with Mental Illness*, n.d.)) prevent people from seeking help (Johnston & Burke, 2020). Stigma can also interfere at an earlier stage, preventing people from interpreting their symptoms as a potential clinical disorder (Schomerus et al., 2019).

1.6. Obstacles to receiving treatment

Timely treatment depends not only on help-seeking, but on a prompt response from the healthcare system (MacDonald et al., 2021).

Care pathway refers to the succession of steps a person who sought help needs to take to obtain treatment, as well as the treatment itself (MacDonald et al., 2018) A study by WHO systematically examined psychiatric care pathways in different countries. The first, most common pathway pattern was identified in countries with the highest number of psychiatrists among the population. In this pathway, most people seek help from a general practitioner, where they are referred to psychiatric services. Other countries have a higher portion of the population self-referring to psychiatric services, or are first referred to a general practitioner by a native or a religious healer (Gater et al., 1991).

The World Health Organization recommended that common mental health disorders (as opposed to severe, long term mental disorders) should be identified and treated by primary care practitioners (Risal, 2011). Some countries do not allow direct access to mental health specialists, requiring a referral from a general practitioner (Wittchen et al., 2003). Unfortunately, primary care physicians are often not perceived as being capable of providing appropriate mental health care. Patients consider general practitioners to be primarily concerned with physical symptoms, and not have the required knowledge and time to address mental health concerns (Fernández et al., 2012; Salaheddin & Mason, 2016; Tunks et al., 2023). Some patients blame dismissive attitude and misdiagnosis by general practitioners as a reason for delays in treatment (MacDonald et al., 2021).

Poor detection of psychiatric disorders among general practitioners was identified before (Daveney et al., 2019; Kondo, 2015, p.; Olariu et al., 2015; Sayal et al., 2002; Vermani et al., 2011). General practitioners have acknowledged the lack of appropriate training to provide mental health treatment (Cullinan et al., 2016). This is complicated by the fact that many patients, especially patients from ethnic minorities, may present with non-specific, somatic (physical) complaints of mental disorders during primary care consultations (Ferenchick et al., 2019; Fernández et al., 2012).

After the initial contact, care pathways can be especially difficult to navigate for young people. Patients and their carers describe their experience as time-consuming and difficult. Some patients report being denied services due to symptoms not being severe enough, not being able to receive treatment until the symptoms escalated to the point of requiring emergency care (MacDonald et al., 2018).

1.7. The promise of universal screening

Noticing and treating symptoms at an early stage can reduce the risk of the symptoms developing into a full-syndrome disorder (De Girolamo et al., 2012). Focusing on prevention and early intervention increases overall wellbeing in communities and reduces the economic cost of mental illness (Farrell & Barrett, 2007). In case of some disorders, it has the potential to prevent a lifetime of disability (Jones, 2013).

Schools provide an unique environment where early-stage symptoms of mental disorders can be identified, and the affected students can be provided with appropriate support, or referred to specialized care (M. Anderson et al., 2019; Bruns et al., 2016; Liao et al., 2023; O'Farrell et al., 2023; Yamaguchi et al., 2020). Many countries offer free and compulsory public education, which ensures that most children who would benefit from an early intervention have an opportunity to receive it, as long as the school is equipped to identify and support them (Bruns et al., 2016). Teachers observe the behavior of their students in a context distinct from that of parents and primary care clinicians. Previous studies have demonstrated that teachers are more effective in predicting future mental health outcomes and help-seeking, at least within the participating schools. (Dwyer et al., 2006; Sharp et al., 2005).

Schools are already a major source of mental health support, especially among rural minority youth (Angold et al., 2002; Mandalia et al., 2018). However, currently most schools take a "wait to fail" approach, where the student is referred for an assessment or specialized services only after they have been struggling academically and/or socially for long periods of time (Bruns et al., 2016; Wood & Ellis, 2022). Universal mental health screening in schools is an alternative approach that has been shown to be feasible and cost-effective in identifying mental disorders in children, especially among the children who are traditionally underdiagnosed, e.g. those in minority groups or rural areas (Connors et al., 2022; Guo & Jhe, 2021; O'Farrell et al., 2023; Wood & Ellis, 2022). Unfortunately, despite encouragement from many government agencies, most schools do not consistently employ universal mental health screening (Connors et al., 2022; Wood & Ellis, 2022), despite being interested in doing so (Wood & McDaniel, 2020).

The most common barriers to implementation of universal mental health screening in schools have been reported to be the lack of awareness about the existence of universal screening programs, no access to screening instruments, and not enough financial resources (Bruhn et al., 2014; Wood & McDaniel, 2020). Additionally, some teachers and school administrators reported a lack of knowledge on how to support students with identified needs, including not

being familiar with referral pathways (Baak et al., 2020; Bruhn et al., 2014). Teachers also report not being confident in their skills of identifying mental health issues in students (M. Anderson et al., 2019; O'Farrell et al., 2023). There is no universally accepted standard for universal mental health screening, and little practical instructions for schools to follow (Connors et al., 2022). School psychologists are not universally trained to select appropriate, psychometrically sound instruments (Siceloff et al., 2017). Systematic evaluation and monitoring of a large number of students is a logistically complex task schools are generally not equipped for (Siceloff et al., 2017). Collecting and scoring screening responses, and keeping track of the students with identified needs can be a time-consuming task to do manually. The software and data infrastructure required to automate those processes, especially if information from several assessments needs to be aggregated, would require a large initial investment which many schools cannot afford (Moore et al., 2015; Siceloff et al., 2017).

1.8. Learning disabilities

Learning disabilities (LDs) are estimated to affect between 5 and 15% of the population (Willcutt et al., 2011). Similar to the mental disorder prevalence numbers, the estimations vary between studies due to differences in definitions and measures (Ozernov-Palchik et al., 2017). The exact definition of learning disability has been controversial, however the underlying concept of "unexpected underachievement" has been a common thread between different definitions. It refers to people struggling to acquire certain skills (e.g. reading, mathematics), despite not having a condition that would interfere with learning of the skill. This condition could be another disorder, or an environmental factor. The terms "learning disorder" and "learning disability" are often used interchangeably (Kronenberger & Dunn, 2003). There is no consensus on the evaluation process for learning disabilities. Criteria for the presence of learning disability in the context of special education varies depending on local legislature (Dombrowski, 2020). DSM-5 requires a difficulty in one of six defined aspects of learning confirmed through a synthesis of individual history, school reports, and psychoeducational assessment (DSM-5).

Besides traditional LDs, other patterns in cognitive weaknesses can affect learning, such as non-verbal learning disability, where people have trouble learning non-verbal kinds of learning such as patterns and concepts, and processing speed deficit (Braaten et al., 2020; Kronenberger & Dunn, 2003).

Learning disorders are often diagnosed after the child starts falling behind their classmates, long after the first signs of difficulties become visible (Gaab & Petscher, 2022; Sanfilippo et al., 2020). This delay in identification and intervention are associated with worse outcomes and development of comorbidities (MacDonald et al., 2018; Mugnaini et al., 2009; Ricky et al., 2017; Vaughn et al., 2007).

Generally, a full assessment for learning disabilities requires administration of cognitive and achievement tests administered by a trained professional. This can take several hours, and be prohibitively expensive (Hayes et al., 2018; Kronenberger & Dunn, 2003; Willcutt et al., 2011). Screening instruments are used to identify children who are at risk and would benefit from further assessment, before they experience significant academic setbacks (Sanfilippo et al., 2020).

Despite multiple countries' authorities recommending or mandating early universal screening for learning disabilities, the recommendations are often not implemented, especially in disadvantaged communities (Schelbe et al., 2022), possibly in part due to lower availability of appropriate screening instruments (Gaab et al., n.d.). The existing screening instruments are often expensive, require a trained professional to administer, and few have confirmed diagnostic validity. DIBELS, the most common screening instrument for reading problems, showed only moderate predictive validity (E. S. Johnson et al., 2009). DIBELS and other commonly used screening assessments rely on administering and scoring brief achievement tests (E. S. Johnson et al., 2009; Soares et al., 2018). An example of a validated question-based screener for learning disabilities is The Colorado Learning Difficulties Questionnaire (CLDQ) (Willcutt et al., 2011).

1.9. Measuring mental health

1.9.1. Properties of psychological tests

In general, measurement can be defined as assigning numbers to objects according to rules. The rules transform qualities of the objects into numbers. Scale is an important quality of measurement rules. For example, to measure heights the scale of centimeters can be used (Kaplan & Saccuzzo, 2001).

A good measure has two important qualities: 1) it measures what it purports to measure, and 2) it produces consistent results between different measurements, given that the measured construct did not change. The first quality is referred to as the validity of the measure, and the

second as reliability of the measure (Gleason et al., 2010). In psychology research, the words *measure* and *test* are often used interchangeably.

There are two main theoretical approaches to construction and evaluation of psychological measurements: most commonly used Classical Test Theory, and the newer Item Response Theory.

In classical test theory, the observed test scores are composed of the true score and measurement error (Magno, 2009). The true score would be obtained if the measurement is perfect. The true score is not the same thing as the underlying construct score. The true score is the expected value of a test (the mean of a large number of independent measurements), and therefore exists only in relation to the test in question. This expected value can be a reliable, but invalid measure of the construct. Validity in CTT is defined as correlations between the true score and some external criterion. (Borsboom, 2005; Cappelleri et al., 2014).

IRT, on the other hand, explicitly represents the underlying construct as a part of the model. In statistics, variables that cannot be measured directly are referred to as latent *variables* (as opposed to *observable variables* that can be measured directly) (Cappelleri et al., 2014; Reise & Rodriguez, 2016). In IRT, responses to items on a test are assumed to be caused by a common, underlying latent variable. Therefore, responses to the items indicate the *location* of the examinee on the latent variable (Reise & Rodriguez, 2016). In the context of latent variables, '*level*', *'position*', or '*location*' refers to the position of an individual along the unobservable trait being measured. It indicates the degree to which an individual possesses the underlying characteristic or trait represented by the latent variable. Higher locations on the latent variable suggest higher levels of the trait being measured, while lower locations indicate lower levels. In the case of educational testing, for example, a higher location on the latent variable might indicate a greater proficiency in the participants being tested, whereas a lower location could suggest a lower level of proficiency (Cappelleri et al., 2014).

Two properties of test items are important to introduce to present the improvement of IRT over CTT: *item difficulty* and *item discrimination*. In CTT, item difficulty is defined as the percentage of respondents who answered the item "correctly" in a given sample. In the context of mental health, the term *item severity* is often used instead – and it is defined by the percentage of respondents who *endorsed* the item. Endorsing an item means choosing an affirmative response to what is being asked in the item (Krishnan & Krishnan, n.d.). Item discrimination refers to the ability of an item to distinguish between high and low scoring individuals (or, for example, individuals with or without a disorder). In CTT both of these

properties are calculated for a particular sample, and are only useful when assessing very similar samples – they are *sample dependent*. The goal of IRT is to obtain *sample independent* item difficulties and item discrimination values (Eleje et al., n.d.).

In IRT, item difficulty defines how much the item contributes to measurement of the underlying construct (the position on the latent variable). The probability of endorsing an item is assumed to monotonically increase as the trait level increases (*dominance response process*). Monotonic increase means that the function does not decrease at any point (although it can stagnate). An advantage of IRT is that any individual item response, or a subset of responses can be used to estimate the position of the individual on the latent variable. This property is widely used in the context of education, in the form of *computerized adaptive testing* (CAT) (Cappelleri et al., 2014; Reise & Rodriguez, 2016). In adaptive (or "tailored") tests items are chosen based on the responses to the previous items, in order to maximize the accuracy and minimize the length of the test. For example, if an examinee correctly answers an item with intermediate difficulty, a more difficult item will be administered. Otherwise, they will be presented with an easier question. CAT has been shown to reduce the number of items in half without any loss in accuracy (Reise & Rodriguez, 2016).

Despite the attractiveness of IRT models, they have been developed mainly with educational testing in mind, their adoption has been slow in the field of mental health (Reise & Waller, 2009). One of the possible stumbling blocks of applying IRT models in mental health assessment is the IRT's requirement for large item pools. It is easier to come up with a large number of items assessing vocabulary than assessing symptoms of depression (Reise & Rodriguez, 2016). Using IRT for test development also requires larger sample sizes (Cappelleri et al., 2014).

This chapter will focus on the CTT approach to test construction and evaluation due to its widespread use. IRT approaches will be further discussed in Chapter 4 (Discussion).

1.9.2. Reliability

There are four ways of assessing whether a test produces consistent results across replications of a testing procedure (American Educational Research Association et al., 2014; Geisinger et al., 2013).

Test-retest reliability refers to the similarity between the scores between two administrations of a test. Considering that the underlying construct remains the same between the two

administrations, the test scores are expected to also remain the same. Test-retest reliability coefficients can be underestimated, because measures of psychological constructs can often be affected by external factors, such as physical discomfort. If a measure is administered by an observer, the observer's perception can change as well. On the other hand, test-retest reliability can be overestimated if the raters remember their responses from the previous administration (American Educational Research Association et al., 2014; Geisinger et al., 2013)..

In *alternate-form reliability* estimation, two non-overlapping versions of a test are constructed and administered to the same participants. For example, for a test measuring vocabulary, two versions can be constructed, each sampling 20 different words out of a full list of 100 words. The test is considered reliable if the scores obtained on the two versions of the test are similar. This method is not often used due to the cost of constructing and administering two different versions of the same test (American Educational Research Association et al., 2014; Geisinger et al., 2013).

The *internal consistency* method is similar to the alternate-form reliability method. It examines how similar responses to different parts of the same test are. If a test measuring one construct has multiple items, the examinee should get a similar score on each half of the test. Using the above example of vocabulary testing, the performance on a subset of 10 words from the full set of words should be similar to the performance on another subset of 10 words, considering they have matching difficulty. If the items have a specific order, for example they are ordered by difficulty, the items can be split into groups of odd and even items (American Educational Research Association et al., 2014; Geisinger et al., 2013).

A generalized version of this process involves comparing responses between each pair of items in the test. If responses are similar between all items, it indicates that all items are indeed measuring the same construct. On the other hand, if the responses are too similar, it might indicate that the items are redundant, and the number of items can be reduced. Tests measuring a more general construct (e.g., anxiety) tend to have lower internal consistency than tests measuring a narrower construct (e.g., math anxiety), because the items need to assess different aspects of the constructs, and therefore are less similar to each other. For this reason, a reliable test measuring a wider construct would generally be longer than a reliable test measuring a narrower construct (American Educational Research Association et al., 2014; Geisinger et al., 2013).

Inter-rater reliability can be examined if the test is administered by a rater (as opposed to the examined individual). It examines how much multiple raters agree on the score of the same participant (American Educational Research Association et al., 2014; Geisinger et al., 2013).

1.9.3. Validity

Standards for Educational and Psychological Testing defines different kinds of evidence that can be used to establish a measure's validity, depending on how it is intended to be used (American Educational Research Association et al., 2014; Geisinger et al., 2013). Different classifications of validity types have been in use in the last century. For example, the 1966 version of Standards separated criterion-related, construct-related, and content-related types of validity. Contemporary validity theory, as described in American Educational Research Association et al. (2014), have moved away from such classifications, instead describing five different sources of evidence of a unitary concept of validity, as described below (Geisinger et al., 2013).

The first source of validity evidence, previously referred to as *face validity*, is based on the *content of the measure*. It is a subjective estimation whether the test appears to be a sensible measure of the measured construct. It is usually done by subject matter experts. First, they evaluate the definition of the measured construct ("*construct definition*"). Then they evaluate if the items of the measure are relevant to the construct – all dimensions of the construct should be covered (e.g. all symptoms of depression for a depression measure), and there should not be any irrelevant items ("*construct relevance*"). Then, they evaluate the proportion of items measuring each aspect of the construct, to make sure that each aspect is appropriately represented in the measure ("*construct representation*"). Finally, the procedure of measure construction is evaluated, including scoring and potential bias ("*appropriateness of test construction procedures*"). In educational testing, an additional step called *alignment* is examined, which evaluates how well the test is aligned with mandated educational curricula (American Educational Research Association et al., 2014; Geisinger et al., 2013).

The second source validity evidence is based on the *measure's internal structure*. The internal structure needs to be assessed in order to defend the type of scores the measure provides, and how they should be interpreted (American Educational Research Association et al., 2014; Geisinger et al., 2013).

Usually, during test development, the authors will hypothesize a specific *dimensional structure* of the construct in question. For example, a test developed to assess the diagnosis

of ADHD as it is described in DSM-5 will hypothesize that the underlying construct has two separate dimensions corresponding to the disorder subtypes: inattention and hyperactivity/impulsivity. The assessment of dimensionality then will apply statistical techniques to the responses to the questionnaire to examine if the hypothesized dimensional structure is supported by the patterns in responses. The dimension in this context is defined as a "homogeneous continuum that accounts for variation in examinees' responses to test items" (Geisinger et al., 2013).

Factor analysis is a common statistical method to evaluate dimensionality of a test (Rios & Wells, 2014). In exploratory factor analysis, the goal is to identify underlying constructs (factors), and determine how much each construct influences each item. The influence of underlying constructs on the items are quantified with "factor loadings". For example, an item about "Avoiding tasks requiring mental effort" can have a loading of 0.8 on the inattention factor, and a loading of 0.1 on the hyperactivity/impulsivity factor. Confirmatory factor analysis is used to test a hypothesized factor structure of a test (American Educational Research Association et al., 2014; Geisinger et al., 2013).

If the author of the test has chosen to use item response theory (IRT) to construct the measure, the evidence of the expected response pattern needs to be provided (Geisinger et al., 2013).

The third source of validity evidence is based on *how it is related to other measures of the construct*. This form of evidence is often called *criterion-related*, where *criterion* is the existing measure that the new measure is compared to. The first step of collecting criterion-related evidence is evaluating whether the selected criterion measures are appropriate. The second step is examining whether the pattern of the results of the criterion measure and the new measure conform to what is expected. One of the simple tests of examining such patterns is the Pearson correlation coefficient. High correlation values are expected between tests measuring the same constructs (so called *convergent validity*), while lower values suggest that the tests measure different constructs (*discriminant validity*). The criterion measure can be administered in a similar time-frame as the test (*concurrent validity*), or later in time (*predictive validity*). Predictive validity is relevant if the test is intended to be used for prediction, for example an academic admission test can be evaluated for prediction of future academic success (American Educational Research Association et al., 2014; Geisinger et al., 2013).

The fourth source of validity evidence is based on the *response process*. Here, the behavior of test-takers is observed through focus groups, interviews, test timing, video recordings, eye movement analysis, and analysis of omitted items. The goal of this process is to examine whether the test-takers use the cognitive process intended by the authors of the test (American Educational Research Association et al., 2014; Geisinger et al., 2013).

The final source of validity evidence is based on the *consequences of testing*. Intended positive consequences of the test should be documented and supported. Possible negative consequences should be documented as well. The goal is not to avoid any negative consequences, but to make sure that the positive consequences outweigh the negative. An example of an unintended negative consequence is the harm caused by inappropriate diagnosis and treatment. Authors of the test are responsible for warning the users of the test of the consequences of inappropriate use. This source of evidence does not lie solely on the test authors, but on test users as well (e.g., school districts where the test is administered) (American Educational Research Association et al., 2014; Geisinger et al., 2013).

Additional statistical techniques can be used to evaluate if the validity findings hold true for different subgroups of examinees (e.g. gender, age). This form of evidence is referred to as the evidence of *measure invariance* (American Educational Research Association et al., 2014; Geisinger et al., 2013).

1.9.4. Diagnostic validity of psychological tests

Establishing a correct diagnosis is a prerequisite for appropriate treatment (Balogh et al., 2015; Falkmer et al., 2013). The diagnostic process involves information gathering and clinical reasoning. The information gathering generally includes an interview with the patient, physical examination, and diagnostic testing (Balogh et al., 2015; W. S. Richardson & Wilson, 2015). In the context of mental health, physical examination can include, for example, observing face expressions and level of alertness. Using the information received during the patient interview and physical examination, their past clinical experience and scientific literature, clinicians estimate the probability of the patient having a particular disorder. This estimation is called a *pretest probability*. If the probability is higher than a certain threshold (*test threshold*), a diagnostic test is warranted. Results of diagnostic tests update the probabilities, generating so-called *posttest probabilities*. When the probability of a diagnosis approaches a certain threshold, the clinician initiates treatment (*treatment threshold*). Test and treatment thresholds depend on safety and cost of the test, effectiveness and availability of treatment, and prognosis of the diagnosis in question (consequences of a missed diagnosis) (W. S.

Richardson & Wilson, 2015). Diagnostic tests have parameters that reflect how useful it is in clinical practice that clinicians take into account during the diagnostic process (Bruno, 2011).

One of the most important qualities of a diagnostic test is its diagnostic accuracy – the test's ability to distinguish people with and without the diagnosis. To establish a test's accuracy, patients are split into two groups: one containing patients with the diagnosis (according to the gold standard measure), and those without. The presence or absence of the diagnosis is established by a reference measure (also called a gold standard measure or a criterion). Ideally, there are no other differences between the two groups besides the result of the reference measure (Fischer et al., 2003; Furukawa et al., 2015).

The most commonly used measures of diagnostic test accuracy are sensitivity and specificity (Ranganathan & Aggarwal, 2018). These measures are calculated based on a 2x2 contingency table with the number of patients with and without the disorder (according to the gold standard measure) as columns, and the number of patients with the positive and negative result of the evaluated test as rows (Figure 2) (Fischer et al., 2003).

2 x 2 contingency table		Target disorder (infection)		Totals
		Present	Absent	
Diagnostic test	Positive	**a**	**b**	a+b
	Negative	**c**	**d**	c+d
		a+c	b+d	a+b+c+d

Figure 2: 2x2 contingency table (reproduced from Fischer et al. (2003)).

Sensitivity represents the proportion of patients with the positive result on the evaluated test among those with the disorder. It is calculated by dividing the number of True Positive patients (patients who got a positive test result and do in fact have the disorder) by the total number of patients with the disorder. *Specificity* represents the proportion of patients who got a negative test result among those who do not have the disorder. It is calculated by dividing the number of True Negative patients (patients who got a negative test result and do not have

the disorder) by the total number of patients without the disorder (Fischer et al., 2003). Test with a high sensitivity prioritizes ruling out the diagnosis in question – if sensitivity is 100%, all patients who have the disease will score as positive on the test, so there will be no patients with the disease among those who scored negatively. If the specificity is 100%, all the patients who scored positive will have the disorder (Drobatz, 2009).

These measures assume that both the reference measure and the evaluated test are dichotomous (only take two values). In reality, most diagnostic tests produce continuous values (e.g. blood glucose, questionnaire score). Before calculating sensitivity and specificities the test scores need to be dichotomized – a cutoff needs to be established, that will separate the positive and negative cases (Fischer et al., 2003).

Receiver operating characteristic curve (ROC curve) can be used to establish the cutoff value. An ROC curve is obtained by plotting sensitivity and specificity values at several cutoff points (Fischer et al., 2003, Figure 3). The appropriate tradeoff between sensitivity and specificity depends on the clinical context. In the situation where missing a disorder and not initiating treatment will lead to serious harm, higher sensitivity is more appropriate. On the other hand, if the treatment is dangerous and should be administered only in most certain cases, higher specificity is preferable.

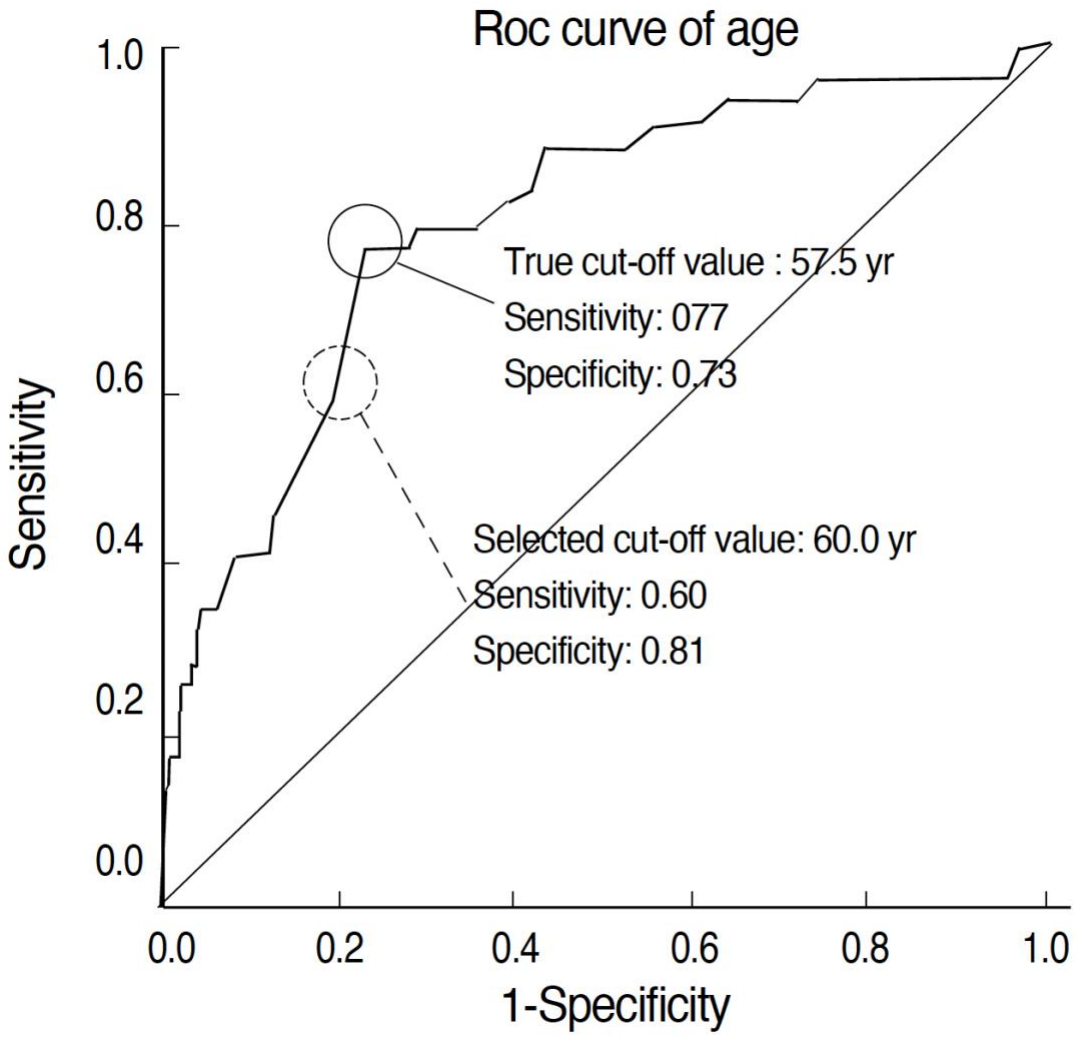

Figure 3: Use of ROC curve to determine the cut-off value of age as a predictor prioritizing higher specificity value ("Selected cut-off value") despite another cutoff resulting in a higher overall performance ("True cut-off value") ") (reproduced from Kwon et al. (2009))

The test cutoff is prevalence dependent, and therefore should be established using a sample where the prevalence of the diagnosis is similar to the prevalence in the population where the test will be administered. The area under the ROC curve (AUROC) represents the overall performance of the test on different thresholds. Since it is independent of the prevalence and of the chosen threshold, it is an appropriate metric to compare two different diagnostic tests (Hasselblad & Hedges, 1995). A diagnostic is generally considered to have moderate accuracy with the AUROC is over 0.7 and high accuracy with the AUROC over 0.9 (Drobatz, 2009).

It is important to take into account how test parameters are evaluated. If a study reporting high test accuracy used only severe cases and completely healthy participants, it is less clinically useful than a test that can identify mild and/or early cases. Additionally, if participants with the diagnosis and participants without diagnosis were recruited from different populations, the accuracy of the test can be overestimated (Fischer et al., 2003; Furukawa et al., 2015). The reference measure against which the test was evaluated needs to be appropriate for every study participant. For example, if an evaluated diagnostic test is a part of a reference measure (e.g. diagnostic criteria used as the reference includes the score on the evaluated test) for all or some participants, the accuracy of the test can be overestimated. This remains true if the evaluated test is not explicitly included in the reference measure, but the person scoring the reference measure is not blind to the result of the diagnostic test in question, and vice versa (Drobatz, 2009; Furukawa et al., 2015).

Sensitivity, specificity, and AUROC values are parts of the criterion-related evidence of validity of a test.

In the context of mental health, diagnostic testing generally involves *screening* and *diagnostic assessments (Balogh et al., 2015)*. Screening instruments identify persons who are considered at risk for a certain diagnosis and are in need of further evaluation. Diagnostic assessments on the other hand are intended to identify the presence or absence of the diagnosis (Charman & Gotham, 2013). Screening instruments generally take the form of a questionnaire or a checklist, either filled out by the patient themself (self-report), or by an observer (parent-report, teacher-report, clinician-report). Diagnostic assessments often take the form of a structured or semi-structured interview administered by a clinician to either the patient or an observer (Chawarska et al., 2008; Kooij, 2012). Usually the goal of both screening and diagnostic assessments is to determine if the diagnostic criteria from a diagnostic manual are met (Kooij, 2012; Regier et al., 1998; Zimmerman et al., 2004).

1.9.5. Screening tools in pediatric mental health

Becker-Haimes et al. (2020) conducted a systematic review of 672 measures to identify brief, free, and accessible youth screening and diagnostic measures across several mental health domains. The measures were evaluated according to their psychometric properties and clinical utility, according to the criteria by (Reyes & Langer, 2018). Measures with excellent psychometric properties as identified by the review are presented in Table 1. No screening measures with excellent psychometric support were identified for suicidality and psychosis.

In addition to the information presented in the review, AUROC and/or sensitivity and specificity was extracted from studies using the original version of the instrument in a general population (primary care or school) children and adolescent sample, with diagnostic interview or clinician-assigned diagnosis as the criterion. Measures for which such studies could not be identified were removed from the list.

Freely accessible screening and diagnostic measures for children and adolescents with excellent psychometric properties

Domain	Assessments	Diagnostic accuracy
Overall Mental Health	Pediatric Symptom Checklist (PSC; Jellinek et al., 1988)	● Any disorder: sens. 0.87/spec. 0.89 (Jellinek et al., 1988) Any disorder: sens. 0.75/spec. 0.75 (Murphy et al., 1992)
	Strength and Difficulties Questionnaire (SDQ; Goodman, 1997)	● Any disorder: sens. 62.6/spec. 86.9/AUROC 75 Individual disorders: avg. sens. 79/spec. 83.1/AUROC 81 (He et al., 2013) ● Individual disorders: avg. AUROC 81.87 (Armitage et al., 2023) ● Groups of disorders: avg. sens. 33.9/spec. 90 (Nielsen et al., 2019) ● Any disorder: sens. 63.3/spec. 94.6 (Goodman et al., 2003) ● Any disorder: 75.5 Groups of disorders: avg. sens. 0.708/spec. 0.788/AUROC 83.8 (Sveen et al., 2013)
Anxiety	Screen for Child Anxiety-Related Emotional Disorders (SCARED; (Birmaher et al., 1997)	● Individual anxiety disorders: avg. sens. 74/spec. 91.48/AUROC 88 (Russell et al., 2013) ● Any anxiety disorder: sens. 79/spec. 82 (Su et al., 2008)

		• Any anxiety disorder: sens. 87.5/spec. 56.1 Individual anxiety disorders: avg. sens. 75.55/spec. 68.2 (Muris et al., 2001)
Depression	Patient Health Questionnaire-9 (PHQ-9; (Kroenke et al., 2001)	• Sens. 89.5/spec. 77.5 (L. P. Richardson et al., 2010) • Sens. 87.1/spec. 79.7/AUROC 0.939 (Ganguly et al., 2013) • Sens. 90/spec. 86.5/AUROC 93.2 (Allgaier et al., 2012)
Disruptive Behavior	IOWA Conners (Loney, 1982)	• Disruptive Behavioral Disturbance: 86/100 (Combination of IOWA Conners and Conners Abbreviated Symptom Questionnaire) (Casat et al., 1999)
	Swanson, Nolan, and Pelham Rating Scale (SNAP-IV; Swanson et al., 2001)	• ODD: AUROC 0.704 ADHD: AUROC 0.877 (Costa et al., 2019)
	Vanderbilt ADHD Diagnostic Teacher Rating Scale (VADTRS; (Wolraich et al., 1998)	• ADHD: sens. 69/spec. 84 (Wolraich et al., 2013) • ADHD: sens. 80/spec. 75 (Bard et al., 2013)

Table 1: Freely accessible screening and diagnostic measures for children and adolescents with excellent psychometric properties identified by Becker-Haimes et al (2020 and their diagnostic validity.

1.8.6. Test development process

Creation of a test starts with the end goal in mind – how are the results of the test going to be used? What information does a score on the test convey? Tests measuring the same construct can look very different depending on the goal and the target audience of the test. For example, test developers need to specify if the test is going to be primarily norm-referenced or criterion-referenced. In norm-referenced tests, the respondent's score is compared to the score of other people who took the test. In criterion-referenced tests the score is compared to some predetermined standard (e.g., a comprehensive diagnostic interview) (American Educational Research Association et al., 2014). After intended use of

the test has been established, the sources of validity evidence described above can guide the test development process (Simms, 2008).

To provide validity evidence based on the content of the measure, the measured construct, the format of the test, and the initial item pool need to be specified. The format specification includes the form of the items (e.g., questions, tasks), form of responses (e.g. multiple choice, free-form), and scoring. Scoring specification includes whether the score will be a simple sum of the item responses, or derived from a more complex scoring model (such as IRT model). The process of test administration needs to be specified as well (e.g., pencil-and-paper vs. software) (American Educational Research Association et al., 2014).

After the purpose of the test, the measured construct, and the test format have all been specified, the test developer assembles a pool of possible items to be included in the test. Items should be written by experts in the field. Item editors who are trained in writing psychometrically sound items can be employed to review the items. The items should be relevant to the specified construct, and representative of all important aspects of the construct. The items should also cover all relevant levels of the measured construct. For example, if authors of a diagnostic test for depression want to be able to assess the whole range of depression severity, the item pool should include both items that discriminate between low and moderate levels of symptoms, and items that discriminate between moderate and severe levels (American Educational Research Association et al., 2014; Geisinger et al., 2013; Simms, 2008).

Common item formats include *dichotomous items* (offering two alternative responses, such as *True or False*) and *polytomous* (offering several alternative responses, such as *Q-sort*, where the respondent is given a list of statements and is asked to give each a rank according to some condition, or *checklists*). *Likert format*, *category format*, and *visual analogue scale* are special cases of polytomous format. In the Likert format, the response options correspond to several levels of agreement with a statement in the question (e.g., strongly disagree, disagree, neutral, agree, strongly agree). *Category format* is similar to the likert format but with more response options. An example of such an item would be "Rate your pain on the scale from 1 to 10". *Visual analogue scale* presents the responder with a 100-millimeter line where a response should be marked between two defined response options.

After an initial item pool is written, the items from the pool are reviewed for clarity, relevance to the measured construct, fairness (e.g., measure invariance described above), and sensitivity issues (potential offensiveness to test takers). More than one person should be

involved in reviewing the items, including persons from historically disadvantaged groups. In the *pretesting* stage the items are administered to a relatively small sample. As an alternative to administering the full item pool, the items can be added to an already administered test. Group discussions can take place at this stage, to capture the cognitive process of test takers when responding to the items (American Educational Research Association et al., 2014; Geisinger et al., 2013; Simms, 2008).

After the item pool is confirmed, it is administered to a representative examinee group, to select a final set of items for the test. The goal of this phase is to gather the validity evidence based on the measure's internal structure. Statistical techniques, such as factor analysis, described in the Validity section are applied on gathered responses. Several rounds of data collection might be required to obtain the final version of the test. Instructions and any required training for test takers are tested out during this stage. If software is used for test administration, security procedures (such as encryption) need to be implemented. Reliability also is evaluated at this stage (American Educational Research Association et al., 2014; Geisinger et al., 2013; Simms, 2008).

The next step of test development is gathering validity evidence based on how it is related to other measures of the construct. Additional validity evidence tests described earlier in the Validity section can be gathered as well. Once sufficient evidence of validity and reliability have been obtained, a publication or a manual should be produced, describing how the test was constructed, how it should be administered and scored, and how the results should be interpreted (American Educational Research Association et al., 2014; Geisinger et al., 2013; Simms, 2008).

1.9.7. Score interpretation

Typically, scoring of a test starts with scores on each individual item. Scores are aggregated into a single score (or sometimes several scores) either by simple addition (sum scores), or using more complicated models (such as IRT models). This aggregated score is called a raw score. Several techniques exist for making raw scores easier to interpret. (Standards 2014, Geisinger)

Sometimes raw scores are converted to scale scores – applying some predefined scaling rules to the raw scores. An example of score scaling is converting raw scores to percentile rank in norm-referenced tests to indicate how the individual compares to other people who took the test. Percentile rank indicates the percentage of people in the sample who got a

lower raw score on the test. To obtain information necessary for scaling a norm-referenced test, the test needs to be administered to a norm group. The norm group needs to be sufficiently large and representative to be valid. For example, scaling rules based on a clinical sample in a large city in the United States will not be valid for calculating scaled scores for a general population in a rural area of a different country. (Standards 2014, Geisinger)

Another approach is defining cut-scores for criterion-referenced tests that separate score ranges into meaningful categories (e.g., no risk/moderate risk/high risk for a disorder). Cut-scores are an equivalent of the cut-off values of diagnostic tests described earlier. Cut-scores can be defined for both raw scores and scale scores. (Standards 2014, Geisinger)

2. Optimizing Item Selection for Psychiatric and Learning Disorder Screening Using Machine Learning

2.1. Introduction

2.1.1. Machine learning overview

Machine learning refers to a set of computational techniques that allow learning complex patterns in the data (Shatte et al., 2019). The most commonly used machine learning method is supervised learning (Nasteski, 2017). In supervised learning a predictive model is fit to the data, that is, it learns the relationship between the input and output variables to be able to predict the output variable for new cases where the output variable is unknown (Nasteski, 2017). During the fitting process model parameters are adjusted to reduce the prediction error (Greener et al., 2022).

Unsupervised machine learning models are able to identify patterns in unlabeled data (Greener et al., 2022), for example to cluster patients with the same disorder into distinct groups based on their symptom profiles. Input variables used for prediction are commonly called features (Greener et al., 2022). Semi-supervised learning is a hybrid of supervised and unsupervised learning, where a part of training examples is labeled (Sarker, 2021). In addition to supervised, unsupervised, and semi-supervised models, reinforcement learning models exist that are trained by reinforcing desirable behavior and punishing undesirable behavior

Before training the model, the dataset needs to be separated into a training set and a test set. The training set will be used to train the model, and the test set will be used for testing its predictive performance (Nasteski, 2017). The individual entries in the training set are called training examples.

2.1.1.1. Supervised learning models

To predict output variables that can take categorical values (e.g. a diagnosis) classification models are used (called classifiers). Models that are used to predict continuous variables are called regression models (Greener et al., 2022). In case of classification, the possible values the output variable can take are called classes. In supervised learning each training example is labeled with class. For example, if a machine learning model is used to predict if a person has a certain diagnosis the two classes could be True and False. Each individual training example will be said to have a label (either True or False). The simplest classification problem

is binary single-label classification. Classification problems with more than two classes are called multi-class classification problems. Classification problems where each training example can have multiple labels are called multi-label classification problems (Dekel & Shamir, 2010).

Supervised learning models are often separated into two groups: linear and nonlinear models. Figure 4 shows two simplified examples of binary classification problems with two features, presented as scatter plots with each training example plotted on a 2-dimensional plane, where each dimension corresponds to a feature. This representation is common, as many machine learning models represent training examples as n-dimensional vectors (2-dimensional vectors in our simplified example). In this example, a linear model will perform well if the two classes can be separated by a straight line (e.g., positive diagnosis if scores on two tests are over a certain value). If the pattern in the data is more complex, a nonlinear model would be appropriate.

Linear models has shown similar performance to more complex nonlinear models in clinical prediction (Christodoulou et al., 2019), possibly suggesting linear relationships between predictors and the outcomes. However some non-linearities have also been reported, for example highest and lowest levels of internalized homophobia among 2SLGBTQ+ youth was associated with the higher risk for a mental health diagnosis during the during the COVID-19 pandemic, while average levels was associated with the lowest risk (Dharma et al., 2022).

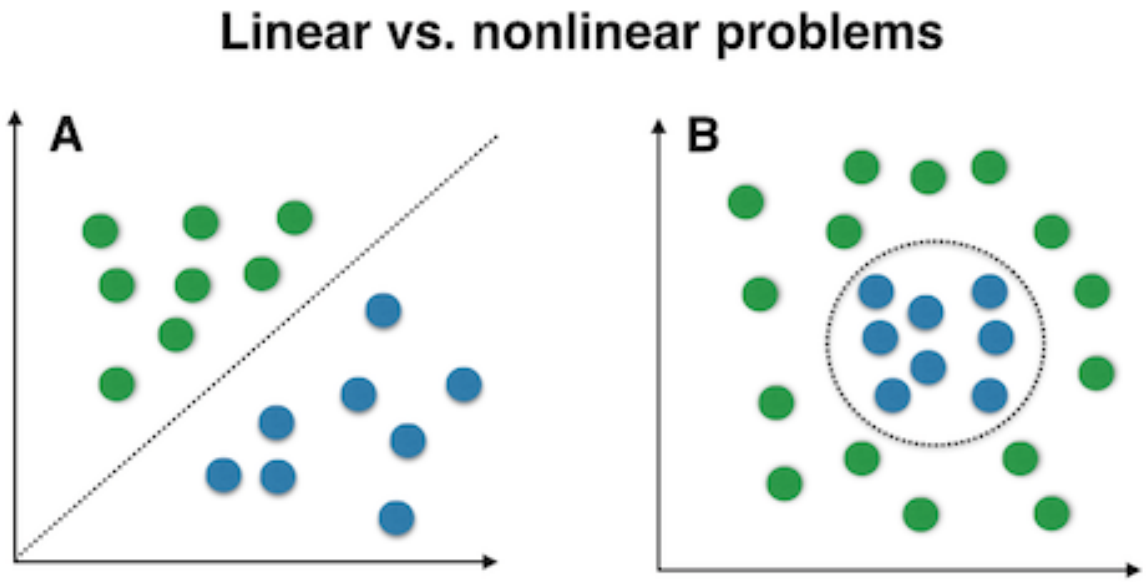

Figure 4: Two simplified examples of binary classification problems with two features, presented as scatter plots with each training example plotted on a 2-dimensional plane, where each dimension corresponds to one feature (reproduced from Activation Functions and Optimizers for Deep Learning Models | Exxact Blog (2019)).

When choosing a machine learning model for the problem other considerations besides the complexity of relationships in the data need to be taken into account. More complex models require larger datasets to achieve good performance on the test set. When applying a complex non-linear model with many parameters to a smaller dataset, a problem known as overfitting can occur. Overfitting happens when a model fits the training data "too well", failing to distinguish relevant patterns from random noise. The opposite of overfitting is underfitting which can happen, for example, when a linear model is fit to a dataset with nonlinear interactions (*Surajustement*, n.d.) . Overfitting and underfitting are illustrated in Figure 5. Generally, performance of linear models stagnates early when increasing the number of examples, while performance of nonlinear models can increase. Model complexity also increases computational resources required for training the model, especially when using a large number of features.

Figure 5: Illustration of overfitting, right fit, and underfitting or a classification model (reproduced from Surajustement (n.d.)).

2.1.1.2. Common machine learning models

Many classification models have been proposed in the literature (Sarker, 2021). A 2009 review on machine learning in mental health showed that the most popular models used for detection and diagnosis of mental disorders were *Support Vector Machines* (SVM), *Regression*, *Random Forest*, *Decision Trees*, *Neural Networks* and *Naive Bayes (Shatte et al., 2019)*. The review showed that most of the models used were classification models, however all the models except NB have both regression and classification versions.

To identify which class an individual belongs to, SVMs identify a *hyperplane* that best separates the classes in the multidimensional space of features (Jiang et al., 2020). Hyperplane refers to a subspace whose dimension is one less than that of the ambient space, e.g. in a 2D space a hyperplane would be a line. SVMs can use a data transformation to support nonlinear class separation. SVM model that does not use any transformation is sometimes referred to as Linear SVM.

Decision Trees assign classes to individuals by building a flowchart where each *node* divides the dataset into groups based on the values of one feature. The *leaf*, or *end node* will contain individuals belonging to the same class (Natarajan et al., 2017).

Random Forest is an *ensemble model* based on DT. *Ensemble Learning* refers to combining several machine learning models into a single algorithm to improve performance. In Random Forest models several decision trees are trained on the data, who vote on the final class label (Natarajan et al., 2017).

Naive Bayesian model is a probabilistic model based on the Bayesian theorem. It requires that input variables are independent from each other, which is unrealistic in most practical

applications. Despite this, it was shown empirically to achieve good results (Viaene et al., 2004).

Neural Networks are inspired by the architecture of the animal brain. They consist of "neurons" organized in layers. The first layer, known as the *input layer*, receives the raw data. Subsequent layers, called *hidden layers*, process information from the previous layer. The final layer, called the *output layer*, produces the final prediction. *Deep learning* refers to using neural networks with many hidden layers, which allows identifying complex patterns in data (Panesar, 2019). Complex neural network models can have millions or even billions or parameters and require substantial computational resources and large datasets (Cheng et al., 2020).

Logistic Regression learns a linear combination of features and maps them to a probability value of the instance belonging to a class (Bartosik & Whittingham, 2021). Similarly to Linear SVM, it assumes a linear boundary between classes. Despite its name Logistic Regression is a classification model – its regression counterpart is Linear Regression (Gudivada et al., 2016).

Since Logistic Regression is based on a linear combination of features, the number of parameters of the model (*coefficients*) depends on the number of input features. Models with a large number of parameters compared to the number of training examples are prone to overfitting. *Regularization* techniques are used to address this problem. Regularization techniques force the learned model to be "simpler" and thus prevent overfitting. LASSO (Tibshirani, 1996) and Ridge (Hoerl & Kennard, 1970) are two regularization methods for Logistic Regression. LASSO "shrinks" lower coefficients to 0, reducing the number of features used in the model. Ridge on the other hand shrinks all coefficients, discouraging the model from fitting the training data too closely. Ridge on the other hand shrinks all coefficients, discouraging the model from fitting the training data too closely (Curtin, 2020; Jiang et al., 2020). *Elastic Net* regularization combines LASSO and Ridge techniques. If a dataset contains features that correlate with each other (and therefore contribute to the output variable to a similar degree) LASSO will choose one coefficient and set the rest to 0. By adding Ridge regularization to the process Elastic Net ensures that groups of correlated features are either retained or dropped out together (Algamal & Lee, 2015).

In clinical prediction literature, Logistic Regression is often categorized as a “traditional” statistical technique, and is excluded from machine learning techniques, which include only non-parametric or non-linear models, such as random forest and neural networks (Christodoulou et al., 2019; Feng et al., 2019; Kuhle et al., 2018; Nusinovici et al., 2020).

Logistic regression and Decision Trees are commonly used when *interpretability* of the classification result is important, however decision trees tend to be unstable (hence the popularity of ensemble models based on combining multiple decision trees), which makes the model less trustworthy (Molnar, 2023).

2.1.2.3. Model evaluation

Classification metrics

There are many evaluation metrics for machine learning models (Panesar, 2019). Performance of classification models is often estimated with Sensitivity, Specificity, and Area under the ROC Curve (AUROC) metrics, described in Chapter 1, and other metrics based on the true positive, true negative, false positive, and false negative rates. These rates are often represented in a *confusion matrix* (Figure 6), similar to the 2x2 contingency tables used in diagnostic test evaluation presented in Chapter 1. (Data Mining Concepts and Techniques). *Classification accuracy* is a commonly reported metric, but it is not appropriate for datasets with imbalanced class distribution. For example, if the prevalence of a disorder in a sample is 1%, and the model simply predicts every instance as negative regardless of the input values, the classification accuracy will be 99%.

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	P'	N'	$P + N$

Figure 6: Confusion matrix (reproduced from Han et al. (2011)).

Some machine learning models, such as Logistic Regression, predict class probabilities, instead of the final classification value. To build a confusion matrix a probability threshold should be established (Oommen et al., 2011). Commonly used probability is 0.5 – if the probability of an instance being of a class is over 0.5, the class label is assigned to the

instance. In binary classification the value of the threshold will affect the balance between sensitivity and specificity – lower threshold will assign more instances to a positive class, thus increasing sensitivity and decreasing specificity. In case of imbalanced datasets, using the 0.5 threshold can result in extreme sensitivity/specificity trade-off with specificity of 100% and sensitivity of 0%. (van den Goorbergh et al., 2022). AUROC is calculated over multiple probability thresholds and therefore independent of the selected probability threshold (Wynants et al., 2019).

The appropriate sensitivity/specificity balance depends on the application of the model, some applications being more tolerant to false positives values, and some to false negatives (van den Goorbergh et al., 2022).

While many machine learning techniques produce a probability estimate, they cannot be used directly in clinical practice. The probabilities are often not calibrated, i.e. they do not correspond to actual probability of the event. Calibration techniques are available to address this problem (Leathart et al., 2017). In case of imbalance datasets, adjusting the probability threshold for class assignment results in better calibrated probabilities compared to correcting imbalance in the dataset before training the model (van den Goorbergh et al., 2022).

Regression metrics

Most commonly used performance metric for regression models is root-mean-squared error (RMSE), or mean absolute error (MAE). They measure the average distance between the predicted and actual values (Botchkarev, 2019; Panesar, 2019). These metrics use the same scale as the output variable, and therefore cannot be used when comparing performance of models predicting variables that use different scales (Botchkarev, 2019). For these cases, the coefficient of determination, or r^2 can be used. For linear models, it explains the proportion of variance in the output variable that is explained by the model. There has been disagreement whether it is valid for non-linear models (e.g., Chicco et al. (2021), Spiess & Neumeyer (2010)).

Cross-validation

An alternative to the simple train/test split k-fold cross-validation can be used, presented in Figure 7. The dataset is split into a defined number (k) of groups, then each fold is used as a test set once, while the rest of the dataset is used as the training set. This way k performance

scores can be obtained, which allows one to examine the stability of the model (how sensitive it is to randomness in the data) (Greener et al., 2022).

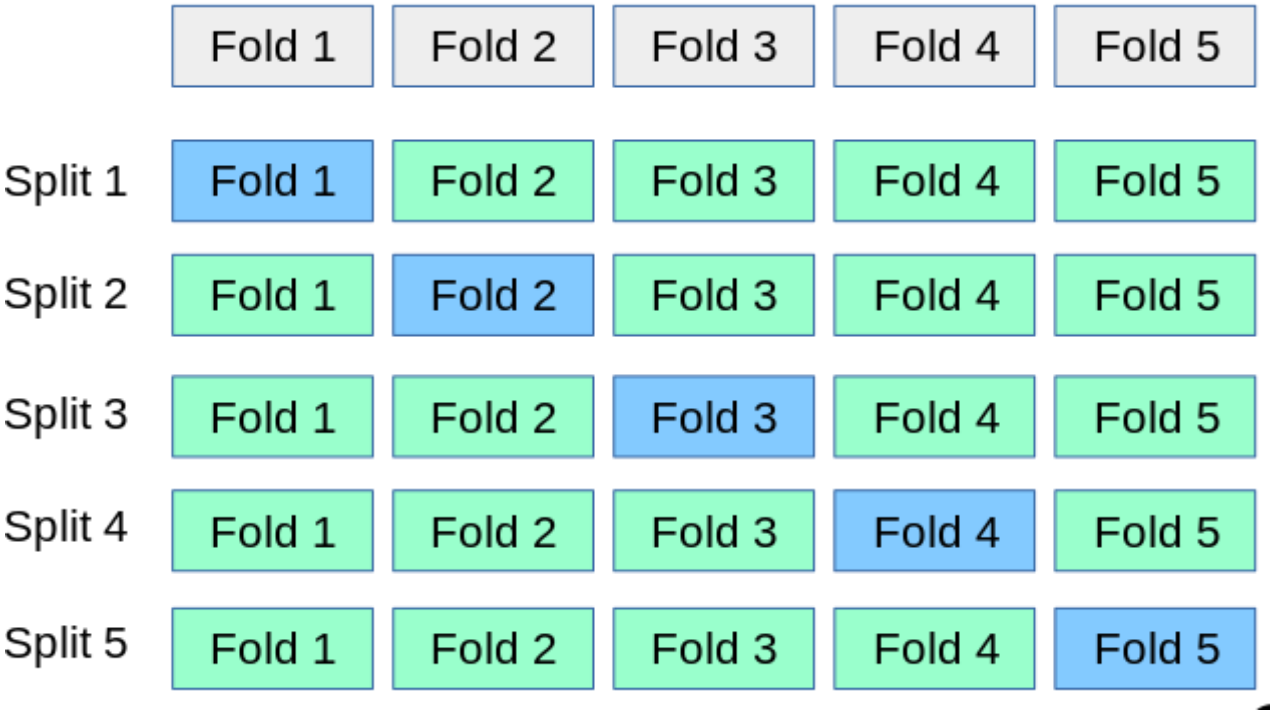

Figure 7: Illustration of k-fold cross-validation (reproduced from 3.1. Cross-Validation (n.d.)).

Other kinds of cross-validation exist, such as leave-one-out cross-validation, where each instance is used as a test-set once, and the rest of the dataset is used as a training set (*Sklearn.Model_selection.LeaveOneOut*, n.d.). Stratified cross-validation matches the class distribution in each fold to the distribution of in the dataset (Han et al., 2011).

Model selection

Some parameters of machine learning models are not learned during the training process but are defined in advance and control the learning process. These parameters are often called hyperparameters of the model. An example of a hyperparameter is the balance between LASSO and Ridge regularization in Elastic Net. It is common to tune hyperparameters by trying out different combinations of hyperparameters and selecting the combination that results in better performance (Raschka, 2020).

Grid Search is one of the common algorithms for hyperparameter tuning. It is an exhaustive search over every combination of predefined values of hyperparameters. A performance

metric needs to be specified for comparing sets of hyperparameters (Panesar, 2019). Each set of hyperparameters is evaluated either using a separate hold-out set or with cross-validation.

Using the test set for estimating the different hyperparameter combinations would result in over-optimistic estimation of model performance. Test set should be used exactly once. A common solution to this is to use a separate validation set besides the train and test set. Yet another alternative is nested cross-validation. Inner cross-validation is used for hyperparameter tuning, and outer cross-validation for estimating performance of the final model (Raschka, 2020).

2.1.2.3. Data preprocessing

Before training machine learning models, the dataset often requires some pre-processing. Data can contain mistakes due to for example typos during data entry. Some data could be missing, either by mistake or by design. Most machine learning algorithms cannot be trained if some values are missing. Common approaches to handling missing data include removing a row with the missing value or removing a feature with many missing values. Missing values can also be imputed – replaced with a value based on the rest of the values in the column, e.g. the average value of the column. More advanced methods exist, for example using a classifier to predict a probable value based on the rest of the columns (Emmanuel et al., 2021).

Some features can be transformed to a more computer-friendly format, e.g. two columns corresponding to the foot part of the height and to the inch part of the height can be merged into one column representing the height in inches. Some machine learning algorithms require that different features have similar value ranges. In this case the values can be normalized to have a similar scale (Panesar, 2019).

2.1.2.4. Feature selection

Feature selection refers to selecting the most informative features out of the whole feature set. Feature selection techniques can be separated into filter methods, wrapper methods, and embedded methods. Filter methods are applied before model training and use properties of the data, for example removing features that are not correlated with the output variable (Miao & Niu, 2016; Venkatesh & Anuradha, 2019). Wrapper methods use the information provided by the trained model, e.g. training the model on several reduced feature subsets and choosing

the subset that results in the highest model performance. Wrapper methods are more computationally expensive but have been shown to achieve better results (Jovic et al., 2015). Embedded methods are a part of the machine learning model itself, such as LASSO regularization for logistic regression (Jovic et al., 2015). Hybrid feature selection methods have also been proposed that combine some combinations of filter, wrapper, and embedded methods (Li et al., 2018).

2.1.2. Feature selection for mental health screening

As described in Chapter 1, the traditional way to create a new screening questionnaire would be to develop a scale consisting of original items written by an expert in the field, and validate it in a target population. An alternative approach would be to use an existing dataset that includes item-level responses to multiple assessments, and computationally identify a subset of items within those assessments that can accurately identify the diagnosis in question.

Feature selection techniques have been used to shorten existing mental health screeners or create new ones. Bone et al. (2016) used a wrapper feature selection algorithm (Forward Feature Selection) with two assessments to improve autism screening. Carpenter et al. (2016) used an embedded method (ADTree) to predict anxiety disorders in primary care using items from a structured interview. Brodey et al. (2018) used a filter feature selection method (Minimum Redundancy Maximum Relevance) to build a 26-item early psychosis screening from a manually constructed item bank of 124 items. Achenie et al. (2019) tested several filter feature selection methods to reduce the number of items in an autism screener. Y. Liu et al. (2021a), Y. S. Liu et al. (2021b) developed a screening application for tertiary care (specialized care within a hospital) based on multiple assessments using Elastic Net for feature selection and tested it for depression, bipolar disorder, and ADHD screening (Y. S. Liu et al., 2023). Tartarisco et al. (2021) used a wrapper method (Recursive Feature Elimination) to select a reduced item subset from a dimensional measure of autistic traits for toddlers to improve early autism screening. Gibbons et al. (2022) combined items from multiple assessments to create a computerized adaptive diagnostic tool that can differentiate between psychotic disorders, using an embedded feature selection method (extremely randomized trees).

Among wrapper feature selection methods, forward sequential feature selection (Forward SFS) and recursive feature elimination were used (RFE). RFE uses the feature importance from the machine learning model it is applied to, for example logistic regression coefficients. At each iteration it removes the least important feature and re-trains the model on the

remaining features. The output of the process is the ranking of the features from the most to the least important.

Forward SFS works by iteratively adding features to the dataset, one at a time, choosing the one feature that produces the greatest increase in performance. The process ends when the specified number of features is reached. An improvement to the original SFS algorithm has been developed (Pudil et al., 1994), called Sequential Floating Feature Selection (SFFS) that checks at each iteration if removing any feature results in an increase in performance.

Bone et al. (2015) described methodological issues in existing research that uses machine learning for mental health screening, and highlighted the advantage of combining items from multiple assessments for creating a new screening instrument. The methodological issues included insufficient familiarity of the researchers with the clinical domain, only classifying severe cases of the diagnosis by excluding less severe cases, using an inappropriate "gold standard" measure, using instances from the training set to obtain model performance, insufficient number of positive and negative instances in the test set, and not reporting stability of selected feature subsets.

Lu & Petkova (2014) compared different filter and embedded feature selection algorithms with the goal of shortening psychiatric screeners, identifying methods that provide the best performance on psychiatric data. In their simulation analysis LASSO and Elastic Net outperformed other methods.

To my knowledge, there were no studies attempting to use feature selection algorithms for question-based screening for learning disabilities.

2.1.3. Healthy Brain Network dataset

In view of the recent issues raised regarding the categorical classification of mental disorders, the Healthy Brain Network (HBN) study aims to address several problems present in current neuroscience research: 1. Studying disorders in isolation from each other 2. Comparing people with disorders to completely healthy controls rather than to people with other clinical conditions 3. Using inappropriately small sample sizes 3. Using clinical as opposed to community-based samples (L. M. Alexander et al., 2017). To address these issues, the HBN study is creating a biobank consisting of a diverse community-based sample of 10,000 children and adolescents residing in the New York City area, which contains physical

measures such as brain imaging and electroencephalography, as well as a comprehensive item-level psychiatric and learning assessment and demographic variables.

Each participant is assessed by a team of clinicians and assigned one or several diagnoses, based on a diagnostic interview (K-SADS, Kaufman et al., 1997) and assessment scores.

The latest HBN data release contains data on 3,625 participants. Exclusion criteria include severe behavioral or cognitive impairment, acute safety concerns, and some neurological concerns. The full list of exclusion and inclusion criteria can be found in Alexander et al. (2017), Table 1. The dataset contains both item-level responses to the assessments, and total and subscale scores for each scale.

The HBN dataset is an appropriate dataset for data-driven assessment creation. First, it includes item-level responses to more than 50 assessments (over 1000 items) targeted at a broad range of disorders. This provides a large pool of items to select from, and offers an opportunity to test if items assessing seemingly unrelated constructs could improve screening efficiency for some disorders. Testing such a large number of items would be impractical in a traditional instrument development study that includes data collection. Additionally, research studies often exclude patients with comorbidities, reporting the performance of their assessments at differentiating between people with the disorder and healthy controls. In a clinical setting, patients often present with multiple disorders, making it challenging to differentiate between patients who have a particular disorder and those who have another, similar presenting disorder. In the HBN study, most participants who were diagnosed with one disorder have one or multiple comorbidities. This provides an opportunity to evaluate the models in a more realistic setting. The other advantage is that the HBN dataset includes both self-report and parent-report assessments. Combining reports from multiple respondents can provide a more comprehensive view of the symptoms, resulting in a more accurate screener.

2.2. Research aims

To build upon the existing research, and address the need for improvement of mental health and learning disorder screening, I develop and test a reusable, generalizable tool that leverages existing datasets to identify parsimonious subsets of items that can be used for mental health and learning disorder screening.

My goals are to: 1) define and test a standardized process for building mental health screening instruments based on existing datasets, 2) identify item subsets that can be used for screening for a set of disorders from the items used in the HBN dataset.

2.3. Methods

2.3.1. Framework description

I implemented a modular framework for identifying item subsets for mental health screening. The framework consists of two packages: dataset preparation package, and item-recommender package. The item-recommender package includes hyperparameter optimization, feature selection, and model evaluation.

The result of executing the four modules is: the recommended item subset, models trained on all item subsets (a subset of the best 1 item, a subset of the best 2 items, etc.), the performance of the models, and the recommended cut-off for optimal sensitivity/specificity values.

I used the structure proposed by the Cookiecutter Data Science package (*Drivendata/Cookiecutter-Data-Science: A Logical, Reasonably Standardized, but Flexible Project Structure for Doing and Sharing Data Science Work.*, n.d.), a standardized project structure for data science projects in python. I changed the proposed structure by moving the trained models to a separate repository, to be able to track changes in the code and in the models separately.

The overview of the framework is presented in Figure 8.

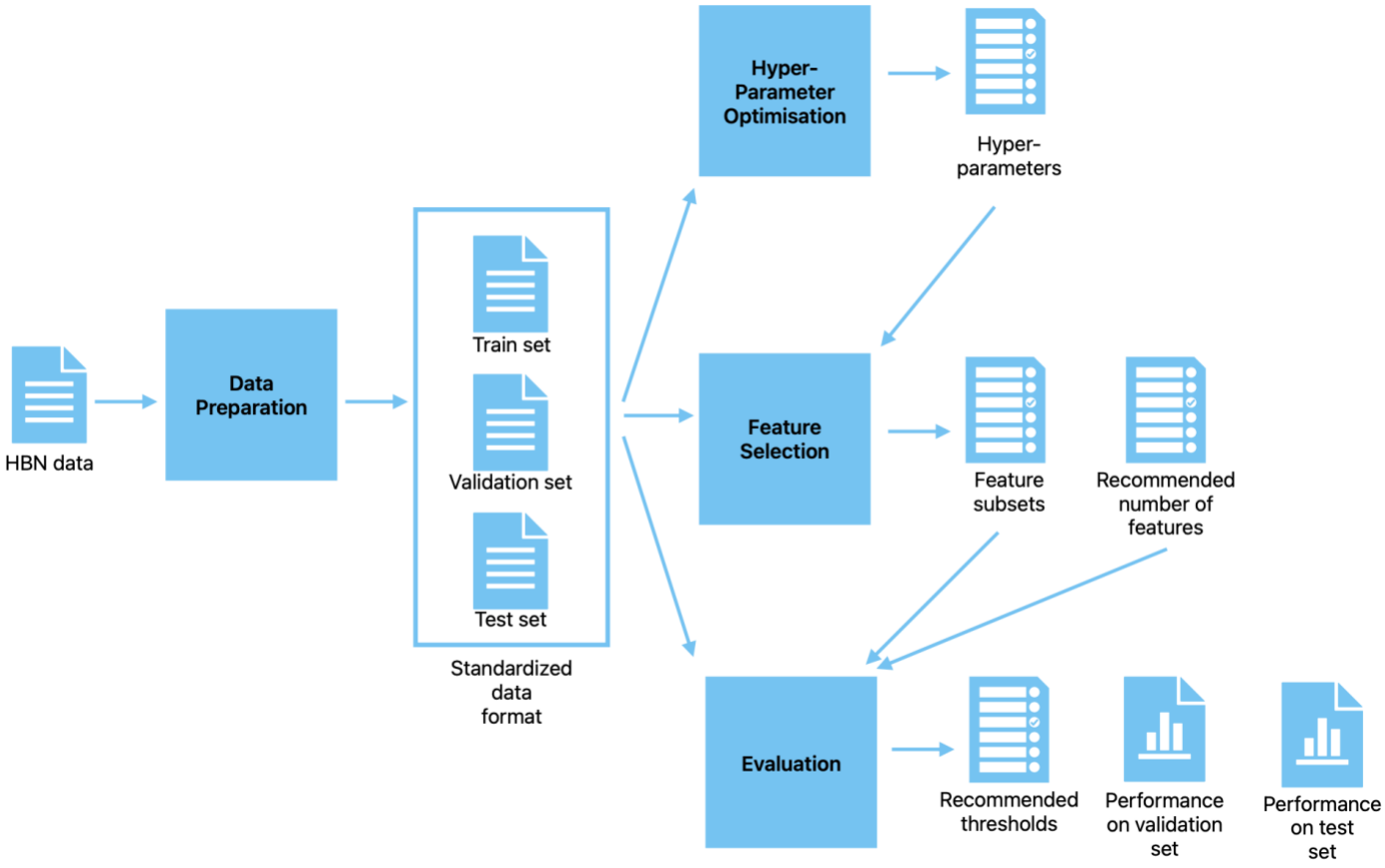

Figure 8: Overview of the framework.

2.3.1.1. Dataset preparation package

In the dataset preparation package, the raw response data is read, transformed, split into training, validation, and test sets, and input and output variables are defined (e.g. which assessments will be used, and which diagnoses will be predicted). The training set is the part of the dataset on which the models are trained; the validation and the test sets are the holdout sets imitating new participants used to evaluate the performance of the models. The validation set was used during the preliminary analysis stage to compare the performance of different models, imputation methods, feature selection techniques, and other variables that can affect the performance of the classification. The test set is used to estimate the performance of the final models.

To ensure that the performance of the models is reliable performance estimates I only predicted diagnoses with more than 20 positive examples in the validation set.

2.3.1.2. Item-recommender package

Hyperparameter optimization

The hyperparameter optimization process was performed using randomized search with cross validation on the training set generated by the dataset preparation package. AUROC was used as the performance metric.

I used the Elastic Net model, as it achieved good performance in previous machine learning research in the mental health domain (see Introduction), and has an advantage of exposing feature coefficients that can guide the development of scoring rules for new screening instruments built based on the identified item subsets.

Feature selection

The feature selection step generates reduced item subsets.

Feature selection is done with Recursive Feature Elimination (RFE, *sklearn* library (Pedregosa et al., 2011)) and Sequential Forward Floating Selection (SFFS, *mlxtend* library (Raschka, 2018)). Feature selection is done using the training set, and the model with the hyperparameters identified by the hyperparameter optimization step.

Due to the high computational complexity of SFFS, I first used RFE to identify top 27 items, and then used SSFS to identify item subsets among the pre-selected 27 features (best subset of 1 item, best subset of 2 items, etc.). Twenty-seven was chosen as the maximum acceptable length of a screener for a single disorder (number of items in the Autism Spectrum Screening Questionnaire, the longest input assessment for a single disorder).

The recommended number of items is calculated by examining the performance of the model for each number of items, and choosing the number of items that reaches 95% of the maximum AUROC value among the evaluated subsets. (The value was changed to 99% for learning disorders due to lower performance).

Subset evaluation

In the subset evaluation step, the models are re-trained on the item subsets using the training set data, and their performance is evaluated using the test set. For identification of binary

variables (presence or absence of diagnosis) AUROC, as well as sensitivity and specificity at multiple thresholds are reported. The optimal threshold is calculated such that, when possible, the sensitivity is higher than specificity (since in screening false positives are better tolerated than false negatives).

2.3.1.3. Addressing class imbalance

Class imbalance was addressed through three strategies. Firstly, to ensure an adequate number of positive examples for learning the data pattern, I restricted predictions to diagnoses with over 20 positive instances in the validation set. Additionally, class stratification was employed when dividing the dataset into training, validation, and test sets, as well as during cross-validation procedures such as randomized search and SFFS.

For meaningful performance evaluation, metrics such as AUROC, sensitivity, and specificity were used. The minority class was defined as the positive class. To address the influence of diagnosis prevalence on the predicted probabilities on, the probability threshold for binary label assignment was computed for each diagnosis. Sensitivity and specificity were reported for various probability thresholds, allowing users to select an appropriate threshold for their specific application.

In addition, *class_weight* hyperparameter was used in the hyperparameter optimization process, which allows adjusting the cost function of the model such that errors in the positive class cost more than errors in the negative class.

2.3.1.4. Output variables

In addition to consensus diagnoses, I created custom output variables for learning difficulties based on performance on achievement tests, which are independent of the input assessment scores.

Following previous literature (Kramer et al., 2020), the criteria for the learning difficulties was defined as IQ being over the two standard deviations below the mean, and the corresponding achievement test subscale being under one standard deviation below the mean. Wechsler Individual Achievement Test (WIAT; Wechsler, 2009) and Wechsler Intelligence Scale for Children (WISC; Wechsler, 2014) scores were used. Wechsler Individual Achievement Test Spelling scale was used for the writing difficulty instead of the commonly used Written Expression scale since the Written Expression scale was not present in the dataset.

Besides the Specific Learning Disorders (SLDs), I also created test-based variables for Borderline Intellectual Functioning (BIF), Intellectual Disability-Mild (ID), Processing Speed Deficit (PS), and two definitions of Non-Verbal Learning Disability (NVLD and NVLD-no-read). Rules used for each test-based diagnosis are presented in Table 2. The rules for ID, BIF, and NVLD are based on criteria from Hetland et al. (2021), Petterson et al. (2007) and Margolis et al. (2020) respectively.

Test-based diagnosis	**Criteria**
SLD-Reading	WIAT_Word_Stnd < 85 **and** WISC_FSIQ > 70
SLD-Math	WIAT_Num_Stnd < 85 **and** WISC_FSIQ > 70
SLD-Writing	WIAT_Spell_Stnd < 85 **and** WISC_FSIQ > 70
ID	WISC_FSIQ < 70
BIF	WISC_FSIQ < 85 **and** WISC_FSIQ > 70
PS	WISC_PSI < 85
NVLD	Criteria used in Margolis et al. (2020)
NVLD-no-read	Criteria used in Margolis et al. (2020), without the condition of non-impaired reading ability

Table 2: Criteria for test-based diagnoses. WIAT_Word_Stnd: Word Reading Standard Score, WIAT_Num_Stnd: Numerical Operations Standard Score, WIAT_Spell_Stnd: Spelling Standard Score, WISC_FSIQ: Full Scale Sum of Scaled Scores, WISC_PSI: Processing Speed Sum of Scaled Scores.

Table 3 presents the final list of predicted diagnoses.

Diagnosis	Abbreviation used
Major Depressive Disorder	MDD
Autism Spectrum Disorder	ASD
Enuresis	Enuresis
ADHD-Combined Type	ADHD-C
Social Anxiety (Social Phobia)	SAD
Generalized Anxiety Disorder	GAD
Oppositional Defiant Disorder	ODD
Any Diagnosis	Any
No Diagnosis Given	None
Separation Anxiety	SA
ADHD-Inattentive Type	ADHD-I
Specific Learning Disorder with Impairment in Mathematics	SLD-Math
Language Disorder	Language
Specific Phobia	Phobia
Specific Learning Disorder with Impairment in Reading	SLD-Reading
Specific Learning Disorder with Impairment in Written Expression	SLD-Writing
Other Specified Anxiety Disorder	Other Anxiety
Processing Speed Deficit (test-based)	PS (test-based)
Borderline Intellectual Functioning (test-based)	BIF (test-based)
Intellectual Disability-Borderline (test-based)	BIF (test-based)
NVLD (test-based)	NVLD (test-based)
NVLD without reading condition (test-based)	NVLD no read (test-based)
Specific Learning Disorder with Impairment in Mathematics (test-based)	SLD-Math (test-based)
Specific Learning Disorder with Impairment in Written Expression (test-based)	SLD-Writing (test-based)

Table 3: List of predicted consensus and test-based diagnoses, and the abbreviations used in this chapter.

2.3.1.5. Input variables

I used all self-report and parent-report assessments except those from the Physical Fitness and Status domain. Due to the nature of the HBN study protocol, most of the assessments were administered only to a subset of participants (i.e. not a single participant was administered all assessments). To minimize the effect of missing data on the model performances, I identified the subset of assessments that was administered to the majority of participants, and only kept the participants who were administered all of the assessments from the subset. This excluded most of the assessments with restricted age ranges such as CBCL/1½-5, which is administered to children under 6 years old.

I also included responses to the pre-intake form containing educational and developmental history of the participant, age, sex, and Barratt Measure of Social Status.

25 most completed assessments were used as the input variables, to limit participants to those who completed all the cognitive batteries required for constructing test-based output variables. Table 4 presents the full list of input assessments.

Assessment Name	**Abbreviation**
Demographics: age, sex	Basic_Demos
Intake Interview - Education History	PreInt_EduHx
Intake Interview - Developmental History	PreInt_DevHx
Child Mind Institute Symptom Checker (*Symptom Checker*, n.d.)	SympChck
Social Communication Questionnaire (RUTTER et al., 2003)	SCQ
Barratt Simplified Measure of Social Status (Barratt, 2012)	Barratt
Autism Spectrum Screening Questionnaire (Ehlers & Gillberg, 1993)	ASSQ
Affective Reactivity Index-Parent (Stringaris et al., 2012)	ARI_P
Strengths and Difficulties Questionnaire (Goodman, 1997)	SDQ
Strengths and Weaknesses Assessment of ADHD and Normal Behavior (J. Swanson et al., 2001)	SWAN
Affective Reactivity Index – Self Report (Stringaris et al., 2012)	ARI_S
Social Responsiveness Scale-2 (Constantino et al., 2003)	SRS
Child Behavior Checklist (Achenbach, 1991)	CBCL

Screen for Child Anxiety Related Disorders - Parent report (Birmaher et al., 1997)	SCARED_P
Inventory of Callous-Unemotional Traits – Parent Report (Essau et al., 2006a)	ICU_P
Alabama Parenting Questionnaire – Parent Report (Essau et al., 2006b)	APQ_P
Parent-Child Internet Addiction Test (Young, 1998)	PCIAT
Distress Tolerance Scale (Simons & Gaher, 2005)	DTS
Extended Strengths and Weaknesses Assessment of Normal Behavior-Parent Report (L. Alexander et al., n.d.)	ESWAN
Mood and Feelings Questionnaire (Angold et al., 1995)ed	MFQ_P
Alabama Parenting Questionnaire (Essau et al., 2006b)	APQ_SR

Table 4: Input assessments and abbreviations used in this chapter.

To check if the trained machine learning models outperform traditional screening instruments, I calculated the AUROC of each total and subscale score of the input assessments for each predicted diagnosis, and compared the AUROCs of the best performing scale to the AUROC of trained model at the same number of features.

Additionally, I checked if the identified item subsets can be scored using simple scoring rules such as sum-scores, without using machine learning models. I used the signs of Elastic Net coefficients to create simple scoring rules: adding together responses to items with positive coefficients, and subtracting responses to items with negative coefficients. I skipped items that had the range of values >6 (such as age) because they would have an overwhelming influence on the prediction, since most of the items have smaller ranges (the majority of the items are rated on the likert scale and have between 2 and 6 possible values).

I repeated the analysis with only non-proprietary assessments, and only parent-report assessments as the input. I used Mann-Whitney test with Bonferroni correction to compare 1) the average AUROC of the best performing existing subscale to the average AUROC of the trained models using all input assessments, and 2) the AUROC of the trained models using all assessments to the AUROC of the models using only non-proprietary assessments, only-parent report assessments, and only non-proprietary parent-report assessments. I performed the identical set of tests on the performances of the sum-scores of identified item subsets.

I investigated the item overlap between the item subsets.

I also investigated if expanding the set of input assessments with less popular assessments, which implies having fewer training examples available, would improve the performance for learning disorders. Table 5 presents the assessments that were added to the input.

Assessment Name	**Abbreviation**
The Columbia Impairment Scale-Parent Report Version (Bird et al., 1993)	CIS_P
Parenting Stress Index Fourth Edition (Abidin, 2012)	PSI
Social Aptitudes Scale (Liddle et al., 2009)	SAS
Repetitive Behavior Scale (Lam & Aman, 2007)	RBS
PhenX Neighborhood Safety (Mujahid et al., 2007)	PhenX_Neighborhood
WHO Disability Assessment Schedule – Self Report (Gold, 2014)	WHODAS_SR
The Columbia Impairment Scale-Self Report Version (Bird et al., 1993)	CIS_SR
Screen for Child Anxiety Related Disorders - Self report (Birmaher et al., 1997)	SCARED_SR
Conners ADHD Rating Scales - Self Report Short Form (Conners, 2008)	C3SR
Children's Coping Strategies Checklist-Revised (Ayers et al., 1989)	CCSC

Table 5: List of assessments added to improve LD prediction and abbreviations used in this chapter.

I also added the scores from the NIH Toolbox – a set of computerized cognitive tasks that can be administered by a teacher or a clinician without any training. The NIH Toolbox scores were excluded from the models for NVLD diagnoses, as it is used in the construction of the diagnosis output variable.

I used Mann-Whitney test with Bonferroni correction to compare the average performance of the original model to the model with the additional assessments, and to the model with the additional assessments the the NIH toolbox scores.

2.3.1.6. Impairment extension

To show how the framework can be applied to different clinical applications, I also applied it to prediction of continuous impairment scores using the same dataset. I predicted total scores of WHO Disability Assessment Schedule and the Columbia Impairment Scale, self and parent reports (WHODAS_P, WHODAS_SR, CIS_P, CIS_SR). The models were changed from their classification versions to corresponding regression versions. Categorical performance measure (AUROC) was changed to a continuous performance measure (r^2).

2.4. Results

2.4.1. Dataset characteristics

The final dataset contained 2,323 participants, with 829 input columns (assessment items).

Age distribution of the whole dataset, and for each consensus diagnosis is presented in Figure 9.

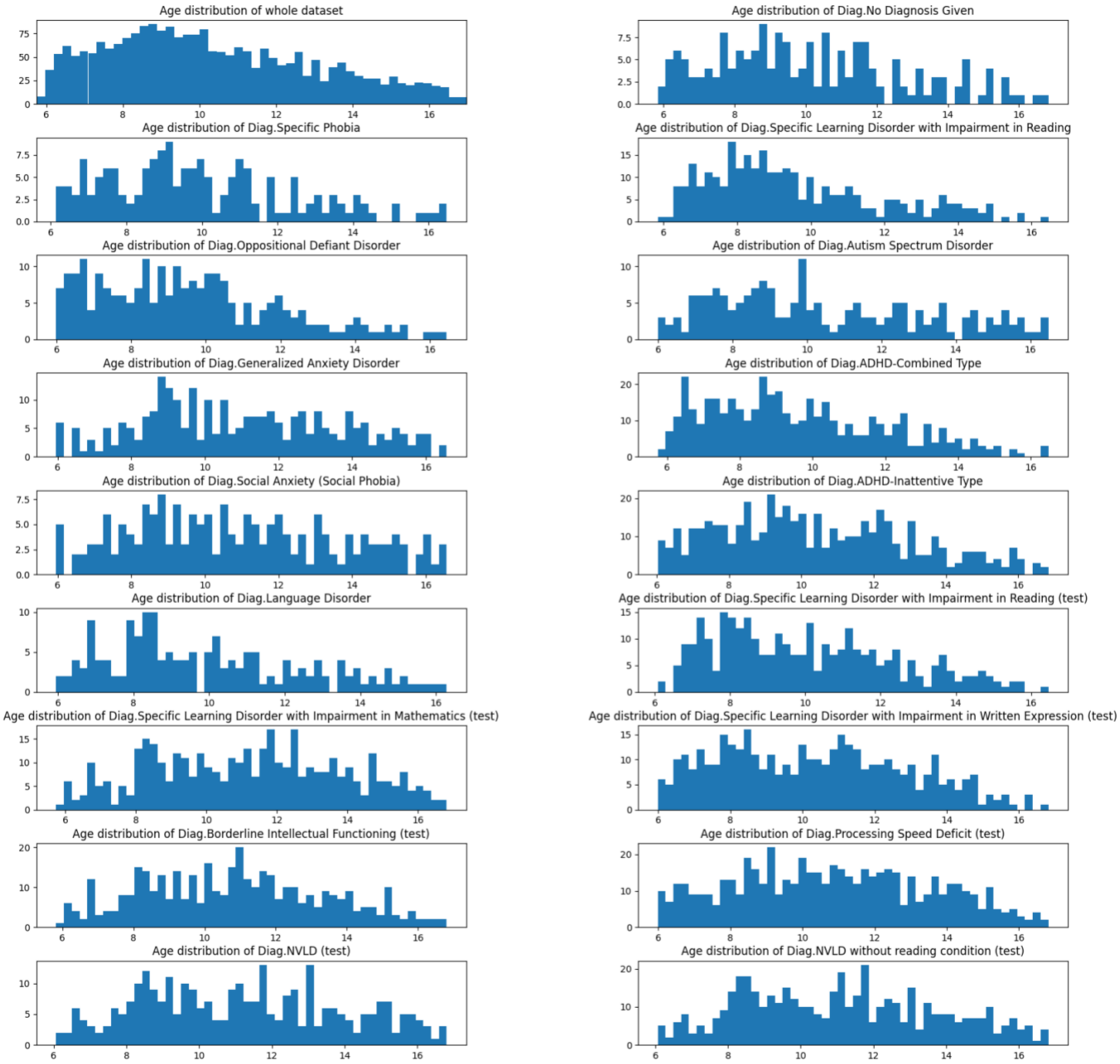

Figure 9: Age distribution of the whole dataset, and per each diagnosis.

2.4.2. Machine learning models vs. standard assessment sum-scores

The comparison of performance of the trained machine learning models and the scores of standard HBN assessments is presented in Figure 10 and 11.

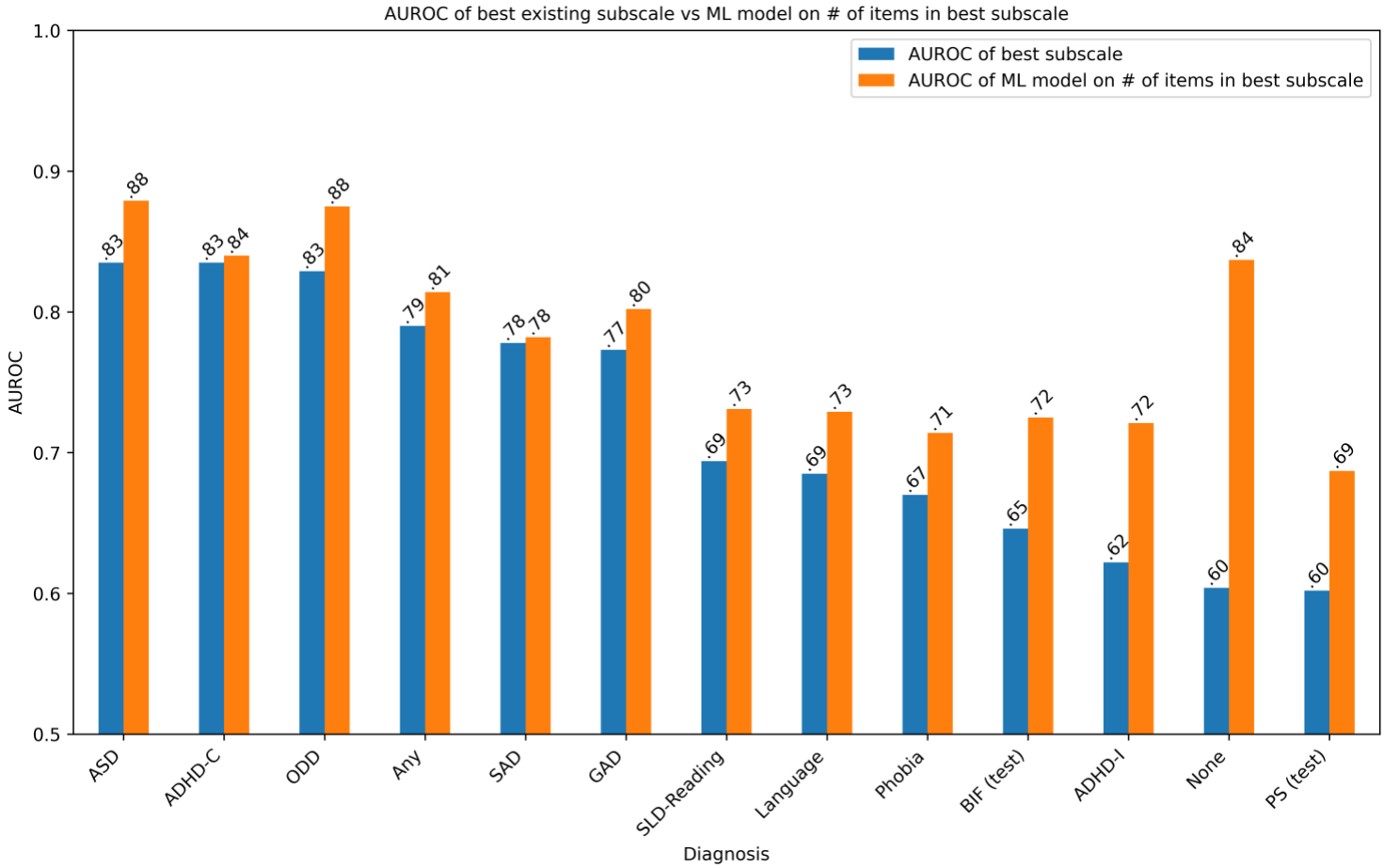

Figure 10: Comparison of the classification performance of machine learning models to the performance of the best subscale (non-learning diagnoses).

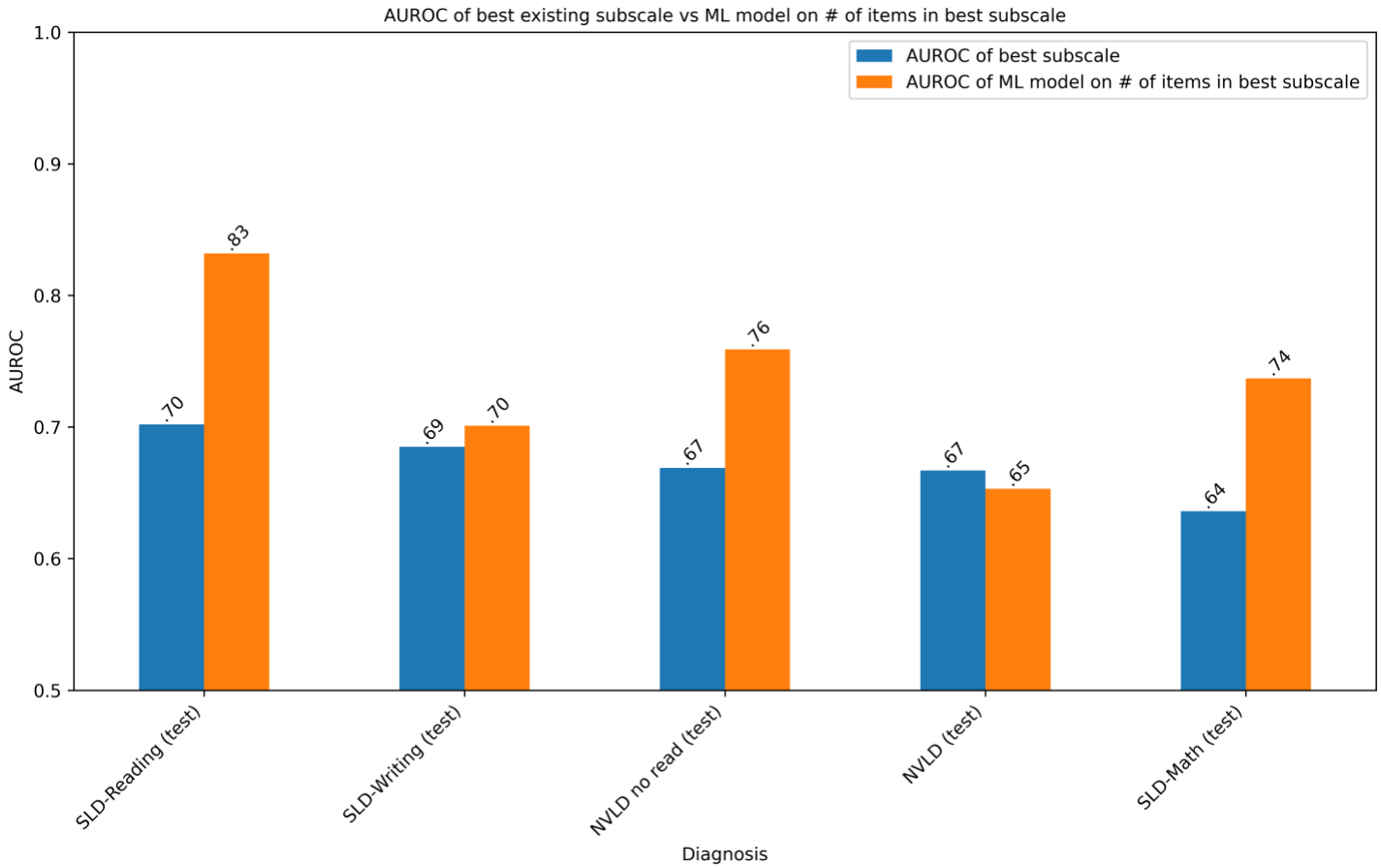

Figure 11: Comparison of the predictive performance of machine learning models to the performance of the best subscale (learning diagnoses).

2.4.3. Subset sum-scores vs standard assessment sum scores

The comparison of the AUROC of the sum-scores derived from learned model coefficients for each diagnosis and the scores of standard HBN assessments are presented in Figures 12 and 13.

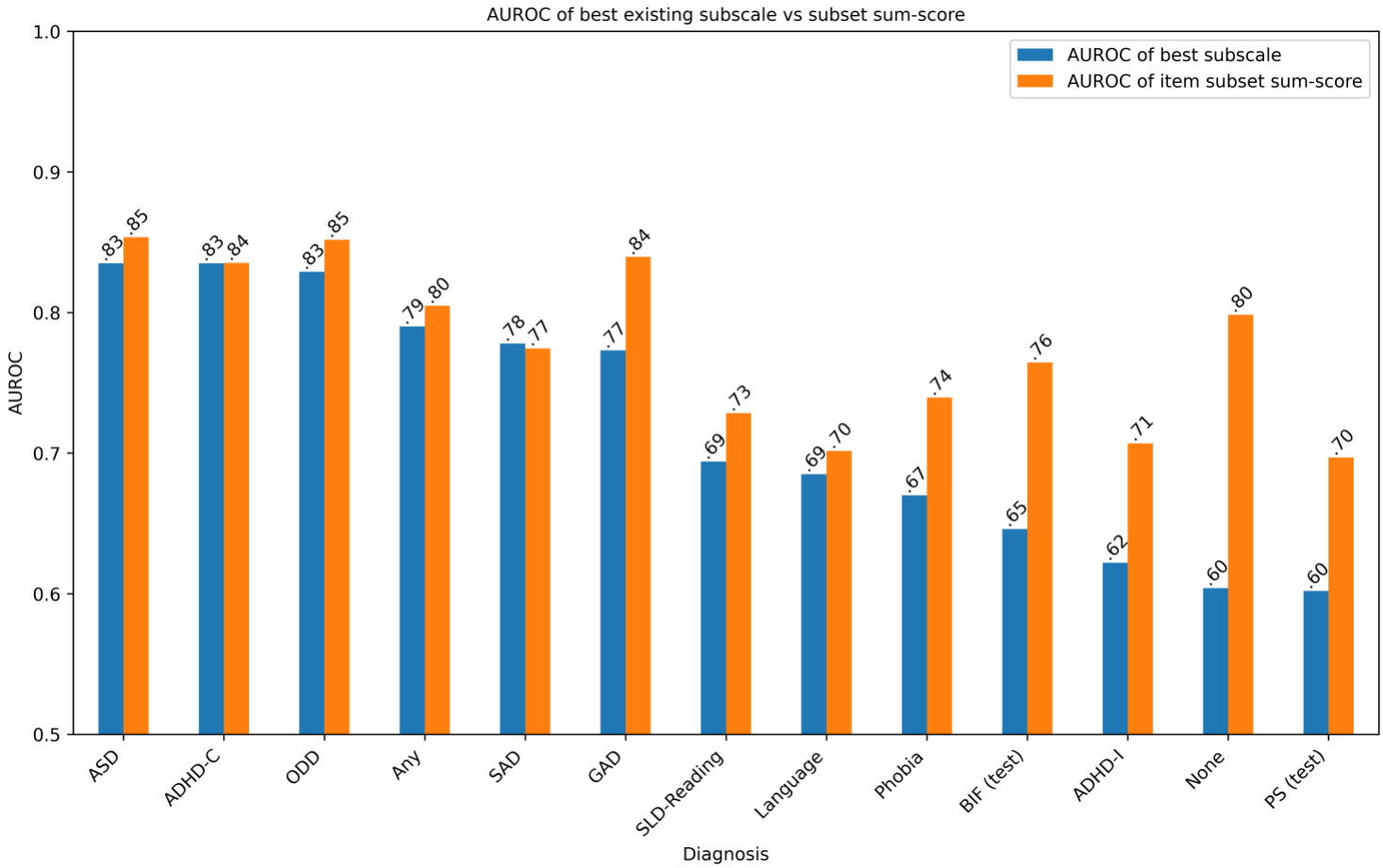

Figure 12: Comparison of the classification performance of the sum-scores of identified item subsets to the performance of the best subscale (non-learning diagnoses).

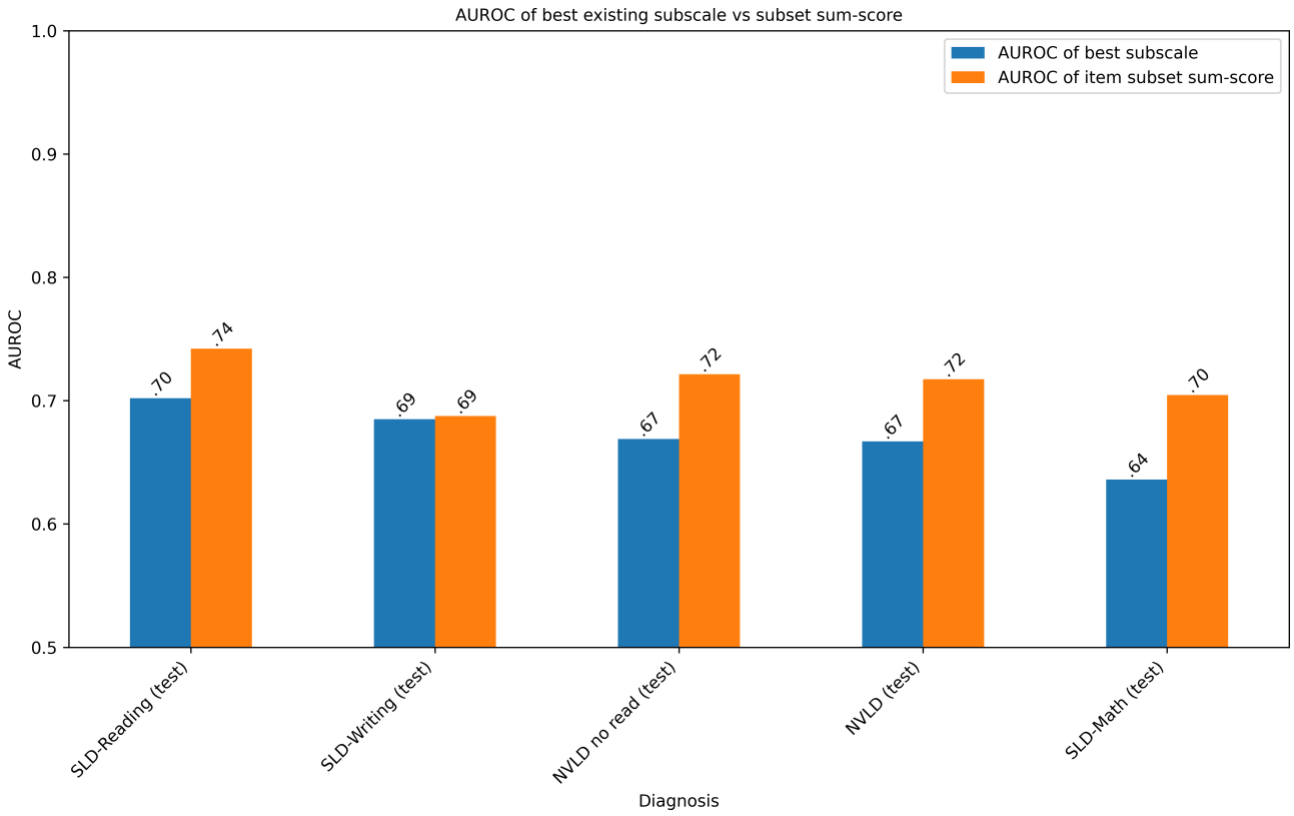

Figure 13: Comparison of the classification performance of the sum-scores of identified item subsets to the performance of the best subscale (non-learning diagnoses).

Figure 14 shows performance of the best existing subscale, the performance of machine learning models using all assessments, only parent-report assessments, only non-proprietary assessment, and only non-proprietary parent report assessment.

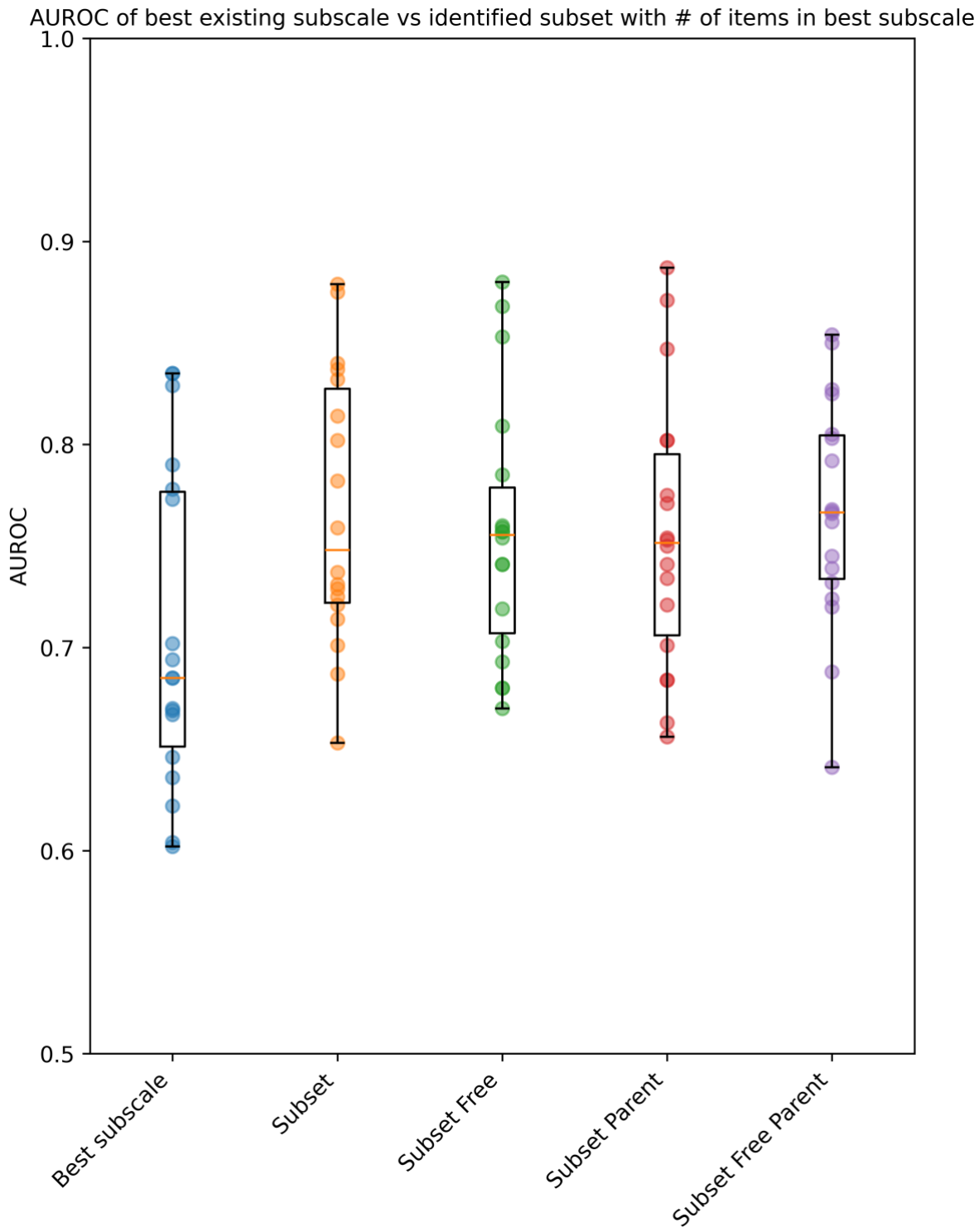

Figure 14: Comparison of the predictive performance of the machine learning models using different input assessment combinations with the performance of the best subscale, averaged between all non-learning and all learning diagnoses.

The Mann-Whitney test with Bonferroni correction showed a significant difference between the performance of the best existing subscale and both the machine learning models and sum-scores of identified item subsets (corrected p-values=0.009, 0.010). No significant difference was found between the performances of machine learning models and sum-scores of identified item subsets using different combinations of input assessments.

2.4.4. Item overlap

Most items were not shared between subsets for different diagnoses. Only five items were in the top five items for more than one diagnosis: *SDQ_02 (Restless, overactive, cannot stay still for long), SDQ_08 (Many worries or often seems worried), SDQ_26 (Overall, do you think that your child has difficulties in one or more of the following areas: emotions, concentration, behavior or being able to get on with other people?), SympChck_05C (Has strong and explosive feelings of anger(Current)),* and *SympChck_51P (Often has a difficult time making eye contact(Past))*.

2.4.5. Improvement of test-based diagnosis scores

Figure 15 presents the difference in performance between the original models, models using additional self- and parent-report questionnaires, and the models using both additional questionnaires and the NIH scores.

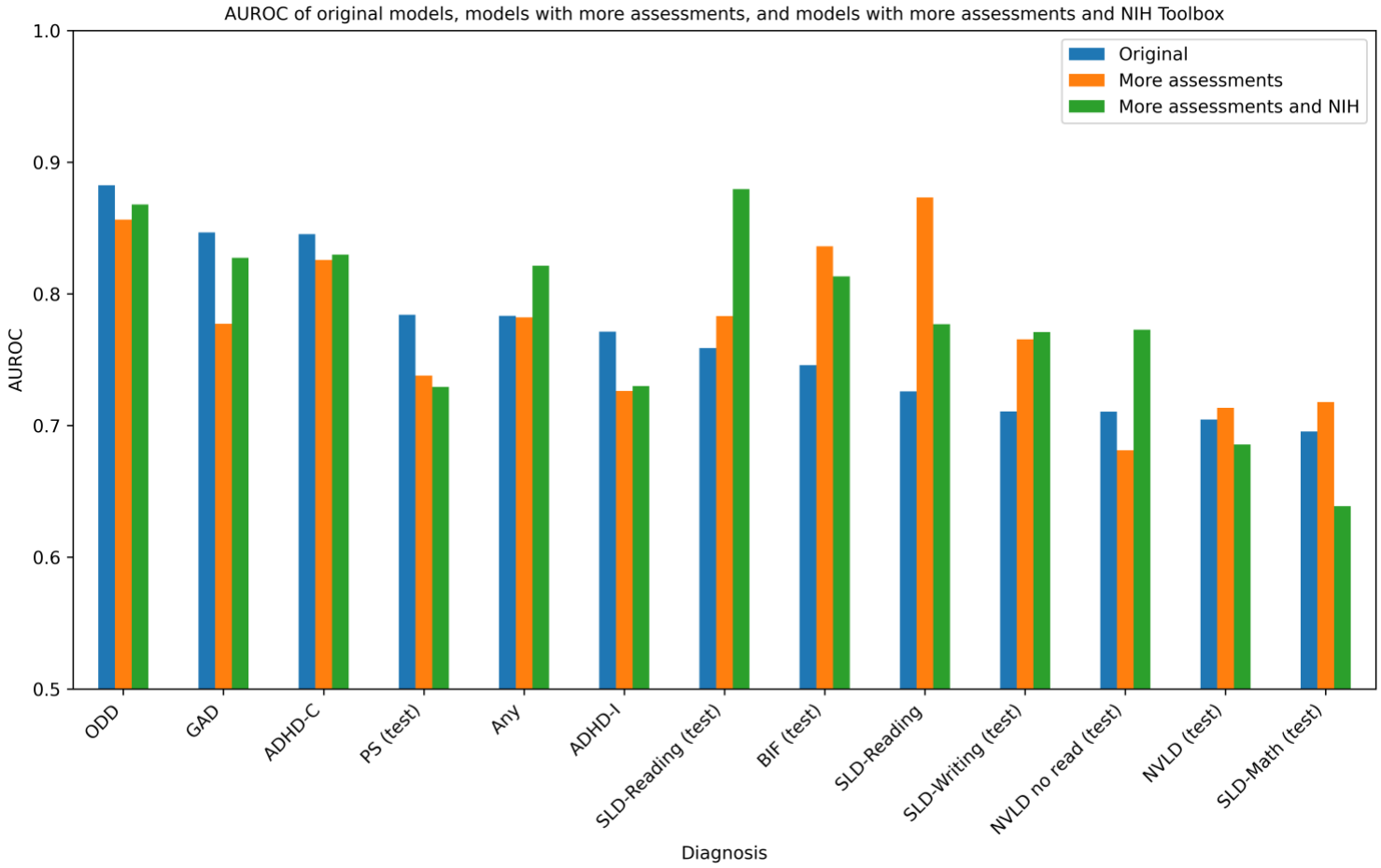

Figure 15: Comparison between the performance of the original models, models using additional self- and parent-report questionnaires, and models using both additional questionnaires and the NIH scores.

The Mann-Whitney test with Bonferroni correction showed no significant difference between the average performances of original models, models using additional self- and parent-report questionnaires, and models using both additional questionnaires and the NIH scores for learning diagnoses.

Top subsets for SLD assessments include items from PSI, C3SR, and NIH Toolbox, which were not among the original input assessments.

2.4.6. Sensitivity and specificity

Tables 6 and 7 present sensitivity and specificity for the identified optimal threshold of the recommended item subsets (non-LD and LD diagnoses respectively).

Diagnosis	Recommended number of items	Sensitivity	Specificity
ASD	6	0.676	0.883
ODD	5	0.851	0.758
GAD	10	0.755	0.687
ADHD-C	6	0.860	0.689
Language	17	0.818	0.502
Phobia	4	0.788	0.389
ADHD-I	11	0.677	0.610
SAD	8	0.784	0.654
BIF (test)	17	0.675	0.711
PS (test)	14	0.780	0.455

Table 6: Sensitivity and specificity for the optimal threshold and recommended number of items for non-LD diagnoses.

Diagnosis	Recommended number of items	Sensitivity	Specificity
SLD-Reading (test)	7	0.811	0.706
SLD-Writing (test)	22	0.851	0.481
NVLD-no-reading (test)	16	0.754	0.593
NVLD (test)	8	0.809	0.488
SLD-Math (test)	9	0.792	0.424

Table 7: Sensitivity and specificity for the optimal threshold and recommended number of items for LD diagnoses.

2.4.7. Impairment extension

Table 8 presents r^2 values for subsets of 27 items for predicted impairment variables.

Impairment score	R^2

WHODAS_P	0.31
WHODAS_SR	0.24
CIS_P	0.69
CIS_SR	0.43

Table 8: r^2 values on subsets of 27 features for functional impairment scores

2.5. Discussion

The goal of this study was to test a standardized process for improving screening for mental and learning disorders using existing datasets. I used the HBN dataset to identify item subsets that can be used for screening for common disorders from the HBN dataset, and to build a machine learning model for SLD screening.

After confirming consent of the authors of the original assessments, the researchers can use the proposed item subsets to construct and validate new screeners, using either simple summary scoring, or scoring using the trained machine learning models (as long as the screened population is similar to the population in the dataset). The performance values of the machine learning models provide preliminary estimates of the performance of the new screener in a population similar to the population in the dataset.

The approach demonstrated improved performance on the HBN dataset for both use-cases over existing assessments, making it a valuable resource for future researchers to build new screening instruments with improved efficacy. The code is available on github.com (Konishcheva, 2022). To apply the method to another dataset it is sufficient to update the dataset preparation package to match the standard data format expected by the item-recommender package.

I showed that the machine learning models trained with my approach produce better performance than any of the existing HBN assessments. Using only items from non-proprietary assessments and only items from parent report assessments did not significantly affect performance, possibly due to the large item pool with many similar items. This is corroborated by the fact that using more assessments did not lead to significant improvement in performance.

The trained model for SLD-Reading showed acceptable performance and could potentially be used for simple, question-based screening in populations similar to HBN, such as schools for children with learning differences. However, the model needs to be validated on the general population before being used for scoring in a more general setting. It is crucial to recalculate the cut-off, since the prevalence is expected to be different in a non-self-referred population.

Since the assessment scores used as the input for the trained models are used by the clinicians to evaluate the consensus diagnoses, the performance of the trained models can be overestimated for non-learning diagnoses. Because of this the performance of these models should only be compared to the performance of other assessments scores present in the dataset. This limitation does not apply to custom output variables based on performance on achievement tests, which are independent of the input assessment scores. Using standardized tests instead of consensus diagnoses also provides a more objective gold standard for the presence/absence of learning disorders, harmonizing the approach with existing research. Another limitation of the HBN dataset for this analysis is that even if the sample is not clinically ascertained, it includes a higher proportion of individuals with clinical symptoms than the general population, due to a recruitment strategy based on perceived clinical concerns.

Another limitation of the study is that the most informative item for all SLDs was from Conner's self-report questionnaire, which is typically administered to children over the age of 8. Generally, it is preferable to screen for learning disorders at an earlier age. Nevertheless, having a screener for older children remains valuable, as not all communities have access to early screening, and many individuals are only diagnosed as adults or not at all. Additionally, the use of items from proprietary assessments like the Conners questionnaire raises potential licensing issues. Non-proprietary instruments that assess learning disability symptoms (e.g. Colorado Learning Difficulties Questionnaire) could be included in transdiagnostic studies in the future. I expect better performance for learning disorders using a dataset containing more items assessing LD symptoms. Compared to the Colorado Learning Difficulties Questionnaire, which includes 6 items evaluating different aspects of reading difficulty, only 1 item in the whole body of items from HBN assessments includes an item directly assessing reading difficulty ("I have trouble with reading").

It is important to acknowledge the limitations of applying machine learning methodologies in the mental health domain. One notable concern is the potential amplification of biases existing in diagnostic practices. Existing bias in diagnosis, such as gender and ethnicity bias in

attributing certain diagnoses, can be exacerbated when using machine learning algorithms, as these models learn from existing diagnostic data. Furthermore, the absence of an objective golden standard test for mental health diagnoses is a fundamental challenge. One of ways to mitigate these limitations is ongoing validation of the models using diverse and representative datasets. The creation of screening tools using machine learning may inadvertently overlook rare symptoms, potentially leading to gaps in the diagnostic process. As machine learning models predominantly learn from prevalent patterns in the data, rare occurrences might not be adequately represented.

Currently, the presented performance scores are obtained from using the trained models to predict the output variables for the examples from the holdout part of the dataset (test set). I am working on obtaining cross-validated performance values, which will provide more robust performance estimates and more stable feature subsets. I will also apply Histogram-based Gradient Boosting Classification Tree (Ke et al., 2017) in addition to Elastic Net, to improve performance of potential screeners that intend to use the trained models as a scoring method. This model showed improved performance for some of the HBN diagnoses in preliminary analysis, but it is not suitable for building screeners scored using sum-scores, since it does not provide coefficients.

Moving forward, I propose exploring multi-label feature selection to identify item subsets for several disorders simultaneously. Additionally, stratifying the training set by age and building separate models for different age ranges could result in more precise screening across different age ranges. This analysis would require a larger dataset, and will become possible with new HBN releases.

To facilitate the application of the trained models as a scoring method, I recommend the development of a user-friendly interface for screener administration and scoring.

2.6. Conclusion

In conclusion, the described framework optimizes item selection for diagnosing common disorders in the HBN dataset, outperforming existing HBN assessments. After confirming validity in the target population and consent from assessment authors, the item subsets can be used for screening of mental and learning disorders. The framework can be applied to other clinical use-cases or new datasets. Future work involves multi-label classification for

simultaneous disorder screening, age-stratified analysis for early identification, and the development of user-friendly screening interface.

3. Learning, Integration, Support, and Awareness Framework

3.1. Introduction

3.1.1. Background

As discussed in Chapter 1, early detection and treatment of mental health and learning disability symptoms can prevent the progression of mental health issues into full-syndrome disorders. Schools offer a unique environment for identifying early-stage symptoms, enabling affected students to receive appropriate support or specialized care. Despite the potential benefits, most schools currently adopt a "wait to fail" approach, referring students for assessment only after prolonged academic or social struggles. Universal mental health screening in schools has proven to be a feasible and cost-effective alternative, especially for underdiagnosed groups such as minorities and rural communities. However, there is a lack of universally accepted standards and practical guidelines for schools to follow, making systematic evaluation and monitoring a complex task for school administrators. Barriers to implementation of universal screening in schools include lack of mental health awareness among teachers and administrators, limited access to screening tools, and insufficient financial resources. The required software infrastructure for screening and monitoring automation entails significant initial investments, posing challenges for many schools.

The Enabee study carried out by the French government in 2022 assessed wellbeing and difficulties of children between the ages of 3 and 11 (*Enabee – étude nationale sur le bien être des enfants*, n.d.). The first results showed that 11% of children were likely to have a mental disorder (Figure 16) – the number close to that identified in the systematic review of prevalence studies across 27 countries discussed in Chapter 1.

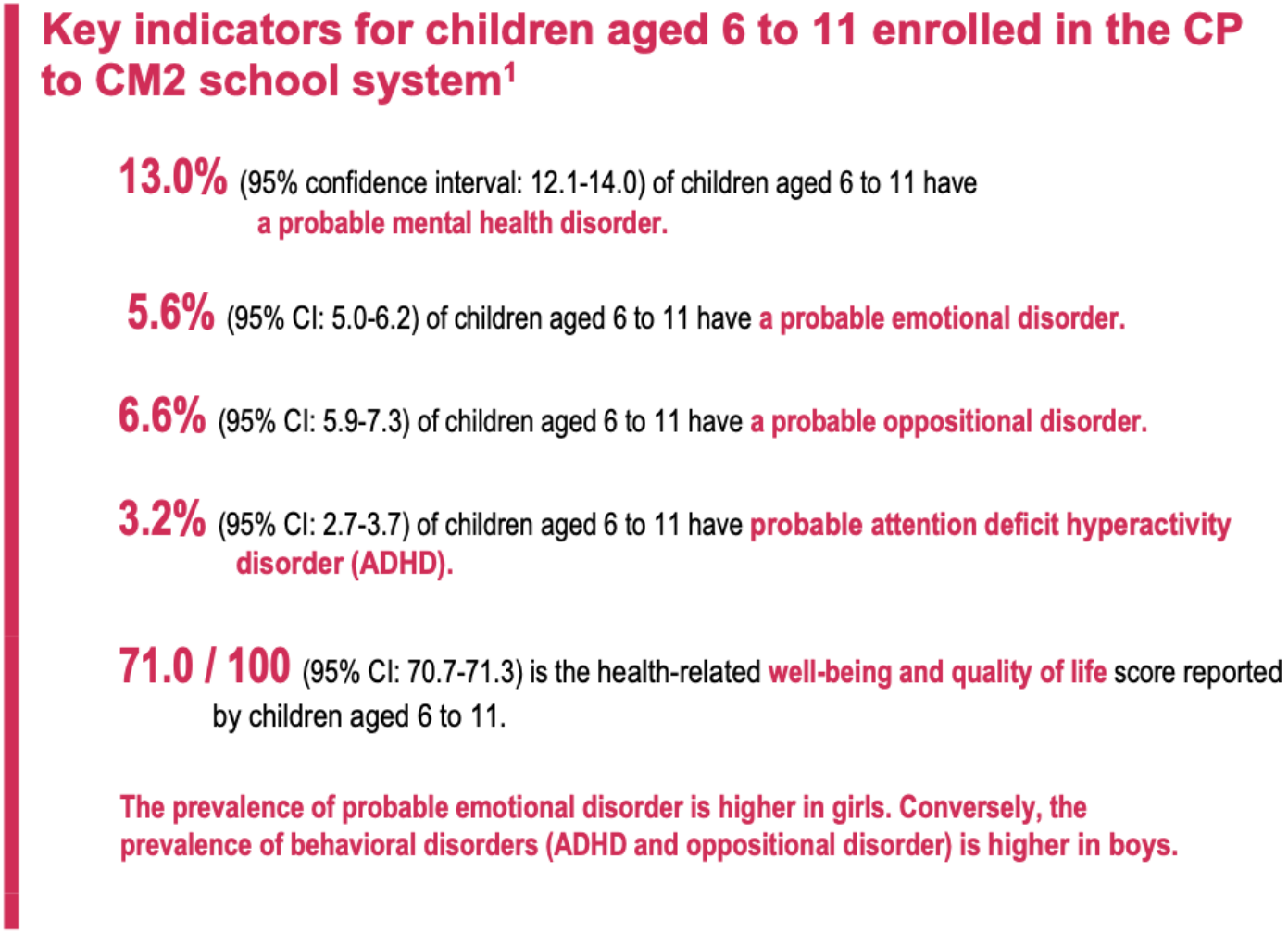

Figure 16: First results of the Enabee study, CP to CM2 refer to first to fifth grade of primary school in the American education system (reproduced from Semaille (2023)).

People with subthreshold symptoms who are not included in prevalence estimates often suffer some degree of functional impairment and are at risk of negative outcomes associated with diagnosable disorders (Institut national de la santé et de la recherche Institut national de la santé et de la recherche, 2002; Polanczyk et al., 2015, 2015; Vaudreuil et al., 2019). Interventions targeted at people with subthreshold symptoms can prevent development of full-syndrome disorder (De Girolamo et al., 2012).

Besides referral to specialized services, teachers are able to provide support to students exhibiting some level of impairment directly in the classroom, such as, for example, regularly communicated expectations, flexibility with timelines, and assistance with planning and organization (C. Johnson et al., 2011; Shelemy et al., 2019). It has been

shown that teachers' support is associated with decrease in problematic behaviors and mental disorder symptoms of students (Shelemy et al., 2019).

Teachers recognize their role in supporting children's mental health needs, but many feel burdened by this responsibility, especially in classes with a higher proportion of students with such problems (Gray et al., 2017). However, the feeling of burden has been shown to be negatively associated with perceived self-efficacy of the teachers in dealing with these issues (Roeser & Midgley, 1997).

Teachers report feeling ill-equipped to deal with children's behavioral and emotional problems and in need of support in recognizing and managing mental health symptoms in their students, and that many children with mental health problems are not overlooked by the school (Forlin & Chambers, 2011; Graham et al., 2011; Gray et al., 2017). Increasing teacher's self-efficacy by providing them with information required to recognize and address problematic behavior can increase their professional commitment and improve students' outcomes (Gibbs & Miller, 2014; Sokal & Sharma, 2013).

A commonly used framework for preventive interventions separates three types of prevention activities: 1. *Universal* preventive interventions that are targeted at the whole population, 2. *Selective* interventions that are targeted at high-risk groups, e.g. children of parents with mental health problems, and 3. Indicated – *targeted* at individuals exhibiting early signs of a disorder (Hagen, 2018)..

Multiple universal school-based interventions have been implemented in recent years in several countries, many showing reduction in mental health symptoms (Fazel et al., 2014). A Cochrane review (Merry et al., 2011) showed that while there is still more support for the effectiveness of targeted intervention, school-based universal interventions have been shown to be effective in addressing mental health symptoms of students in both low- and high-risk groups.

Schools' role in student's well-being has been included in French law in 2013 by establishing the "educational health pathway" (le parcours éducatif de santé). This program aims to structure health education, prevention, and health protection among youth, including early identification of health problems that can affect learning (Labaye-Prévot et al., 2022). As a part of this effort, Health Promoting School (École promotrice de santé), established in 2020, aims to strengthen the coordination of all health-promotion

initiatives, improve environmental conditions at school, and encourage healthy behaviors in pupils by developing prevention from an early age. The Health Promoting School initiative established an Édusanté label system (*Je souhaite m'engager dans la démarche École promotrice de santé*, 2023). Édusanté label certifies a certain level of expertise, and fosters a common culture around health promotion, ultimately encouraging inter-institutional exchanges. The label is available at three different levels, ensuring accessibility for schools, colleges, or high schools involved in health promotion initiatives for students and staff. (*Le Label Édusanté*, n.d.). The initiative also encourages inviting trained external speakers, and provides several training courses aimed at teachers.

An official bulletin issued by the French Ministry of Education in 2016 emphasized the importance of health in its physical, psychological, social, and environmental dimensions. It highlights that health promotion in schools is essential for students' educational success, including health education and prevention projects. The pathway is organized around three axes: health education, prevention, and health protection. The bulletin emphasizes collaboration between schools, establishments, districts, and local authorities to support students' health and well-being (*Mise en place du parcours éducatif de santé pour tous les élèves*, 2016).

3.1.2. Student-focused education support resources in France

The French system for supporting children with learning difficulties offers various options to ensure their education under suitable conditions. These options include personalized educational programs such as the "Programme personnalisé de réussite éducative" (Personalized Educational Success Program, PPRE), "Projet d'accueil individualisé" (Individualized Reception Plan, PAI), "Plan d'accompagnement Personnalisé" (Personalized Support Plan, PAP), and "Projet personnalisé de scolarisation" (Personalized Education Plan, PPS). These programs are designed to provide tailored support, addressing specific learning needs and disabilities. The PPRE focuses on pedagogical support for students struggling with essential knowledge and skills, and it is mandatory in case of grade repetition. The PAI caters to students with chronic health conditions, specifying adaptations and medical treatments. The PAP addresses students with persistent learning difficulties due to learning disorders, offering pedagogical accommodations and support. The PPS is for students officially recognized as having disabilities by the "Commission des droits et de l'autonomie des personnes handicapées" (Commission for the Rights and Autonomy of disabled People) and provides comprehensive educational, psychological, and medical support tailored to the student's needs. These programs aim to ensure an inclusive and quality education for all students

from preschool to high school, considering their individual needs through adapted pedagogical actions. PPRE is initiated by request of the teacher if a child does not meet the expected educational achievement. PAP requires an official learning disorder diagnosis (*École et handicap - PPS, PAI, PAP, PPRE*, 2021).

"Guide d'évaluation des besoins de compensation en matière de scolarisation" (Guide for Evaluating the Needs for Compensation in Education, GEVA-sco) is a standardized assessment tool for evaluating the needs of students with disabilities or special educational needs in the context of their education. GEVA-sco provides a structured way to assess a student's needs, which is then used by a multidisciplinary evaluation team to make decisions related to the student's educational placement, support measures, material adaptations, and educational accommodations. GEVA-sco includes an assessment of difficulties the child faces during the education process, including cognitive, social, mobility, communication, and hygiene support needs (*École et handicap - Qu'est-ce que le GEVA-sco ?*, 2021).

The "Réseaux d'Aides Spécialisées aux Élèves en Difficulté" (Specialist Support Networks for Students in Difficulty, RASED) provides specialized assistance to students facing significant learning challenges in elementary school. Specialized teachers and educational psychologists work alongside regular teachers to address learning and adaptation difficulties experienced by some students. These specialists contribute to the development and implementation of PAP and the monitoring of PPS. The specialized assistance aims to prevent and remediate academic difficulties that persist despite efforts by classroom teachers (*Les réseaux d'aides spécialisées aux élèves en difficulté (Rased)*, 2014).

If the support by PPRE, PAP, PPS, and RASED is insufficient, starting middle school a child can be enrolled in Sections d'enseignement général et professionnel adapté (Adapted General and Vocational Teaching Section, SEGPA). It is an educational structure designed to support students with severe learning difficulties. The children are placed in a small group of students (maximum sixteen) to individualize each student's education (*What Is a Segpa Class?*, 2023).

An equivalent of SEGPA for high school students is the "Établissements régionaux d'enseignement adapté" (Regional Adapted Education Institutions, EREA) and "Lycées d'enseignement adapté" (Adapted Education High Schools, LEA). They are local public educational establishments that cater to students facing significant academic, social, or

disability-related challenges. They offer a unique blend of adapted education, vocational training, and pedagogical and educational support (*Les établissements régionaux d'enseignement adapté*, 2023).

"Le livret de parcours inclusif" (the Inclusive Path Booklet, LPI) is a digital application designed to provide educational solutions for students with special needs. It supports the implementation of personalized educational programs such as PPRE, PAP, PAI, and PPS. Aimed at professionals including teachers, school administrators, and medical professionals, the LPI offers quick and effective implementation of accommodations and adaptations based on a database of resources. It simplifies the procedures for creating and editing plans, allowing collaboration between schools and families. Guides and resources are available for different user roles, including school directors, teachers, and support staff (*Le Livret de Parcours Inclusif (LPI)*, 2023).

3.1.3. Educator-focused education support resources in France

Canopé is a network of regional educational resource centers in France, managed by the French Ministry of Education. Canopé centers provide a wide range of resources, including teaching materials, educational tools, multimedia resources, and pedagogical support to enhance the quality of teaching and learning in schools. Canopé also organizes training sessions, workshops, and conferences for teachers, allowing them to stay updated with the latest educational practices and methods (*Qui sommes-nous*, n.d.). One of their aims is to provide teachers with resources needed to accommodate children with special education needs and disabilities. They provide an online tool that consists of a comprehensive questionnaire of 101 items that teachers can fill out about their student to assess their needs, which includes assessment of different aspects of language and communication skills, cognitive and motor skills, and personal and emotional development (*Étape 1 : je compose ma grille d'observation*, n.d.). Teachers can fill out the whole questionnaire or choose specific areas. An excerpt from the questionnaire is presented in Figure 17.

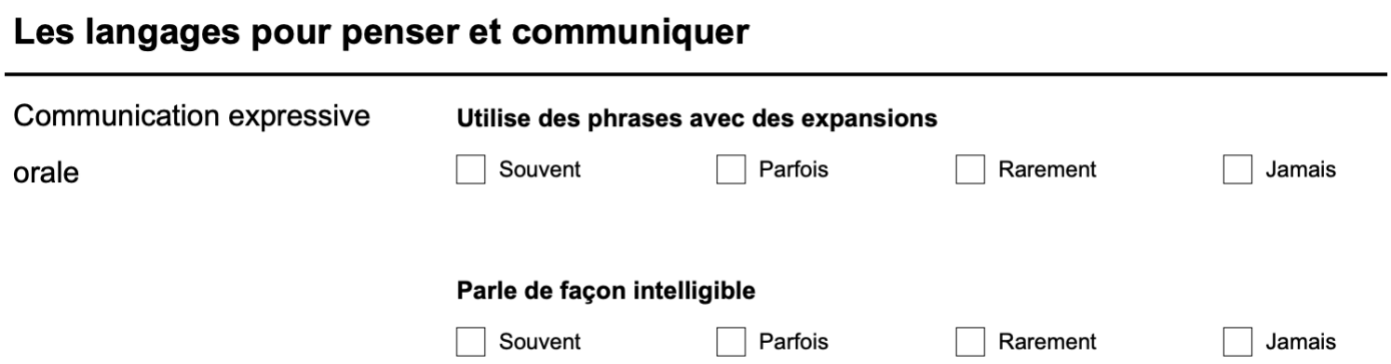

Figure 17: Excerpt from the Canopé assessment of students' needs (Étape 1 : je compose ma grille d'observation, n.d.).

After the teacher fills out the questionnaire, they are shown the identified problems, and a list of strategies to accommodate the student. Each strategy is presented as a one-page guide with a context section describing how the strategy can help the student, and "Adaptation activities" with specific actions the teacher can take to accommodate the student (*Mettre en confiance l'élève pour faciliter sa production orale*, n.d.).

The online platform also contains an information section with a set of guides presenting common mental, cognitive, and physical problems that affect students' wellbeing (*S'informer*, n.d.).

The French Public Health Agency provides a list of evidence-based interventions aimed at health promotion that can be adopted by decision-makers and local stakeholders. Among programs aimed at youth mental wellbeing it includes several programs for adolescents for smoking prevention and alcohol consumption reduction, emotional and social development programs and suicide prevention programs (*Répertoire des interventions efficaces ou prometteuses en prévention et promotion de la santé*, n.d.).

The French Public Health Agency also published a report on the state of scientific knowledge on psychosocial skills in France as of 2021. Psychosocial skills include cognitive, emotional, and social skills. The document identifies key factors common to successful psychosocial skills programs. The main factors of effective programs included formal and structured approach, well-prepared facilitators and an organized CPS team comprising various stakeholders. The report acts as a scientific foundation and is said to be later supplemented by practical resources for stakeholders (SPF, 2022).

Despite the governmental effort to include mental health in the education system, the adoption of such programs in schools remains low. Patalay et al. (2017) surveyed schools in 10 European countries, including 80 schools in France, to assess schools' mental health provisions. French schools reported the lowest levels of interventions in all categories of mental health programs, including both whole-school and targeted interventions (Figure 18).

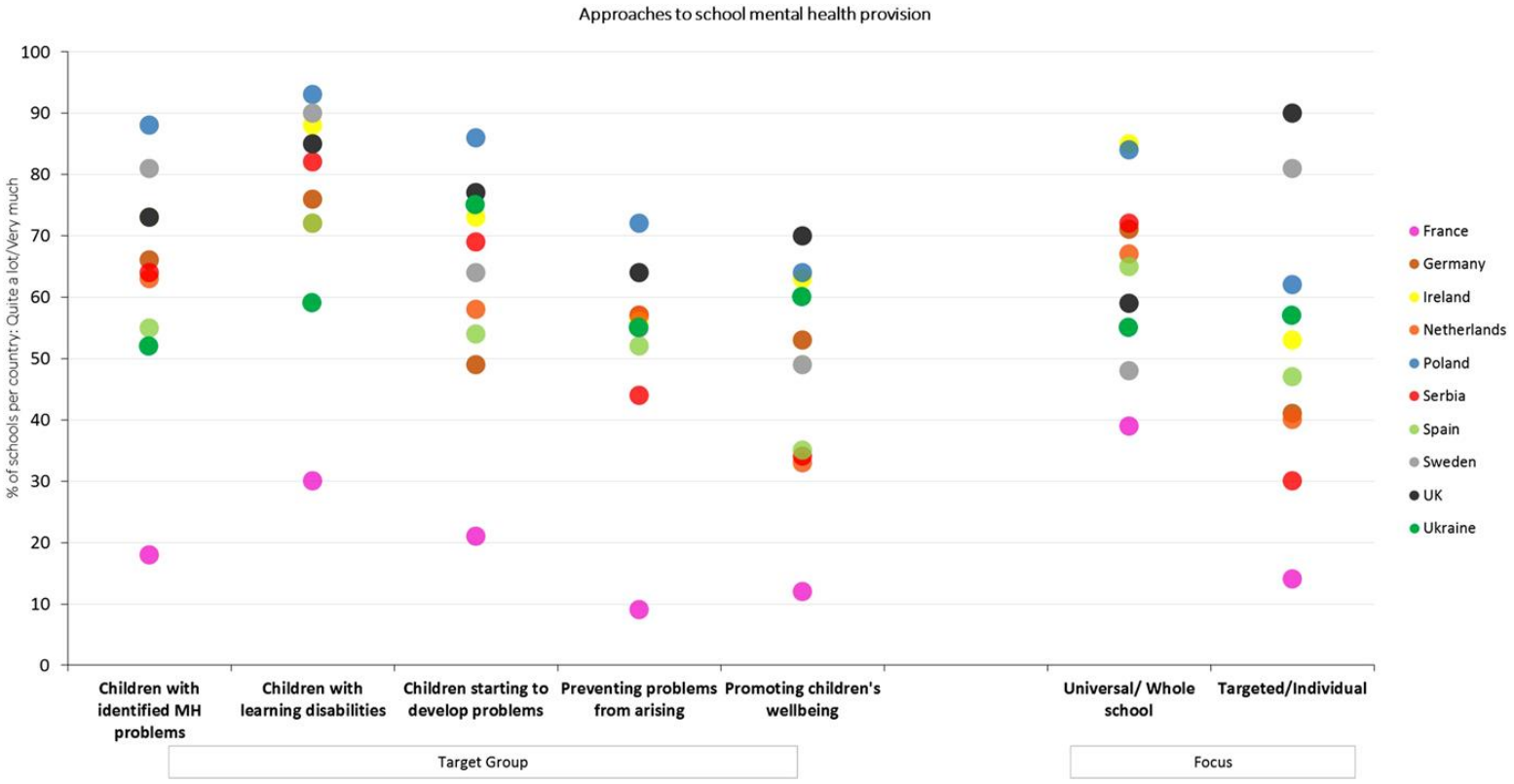

Figure 18: Percentage of schools reporting available mental health provision of different categories (reproduced from Patalay et al. (2017))

3.2. Challenge and objectives

The challenges identified in the current landscape of mental health management of school-age children involves a range of systemic issues. The first is the low proportion of children receiving the necessary treatment, in comparison to the estimated prevalence of mental health problems. Similarly, personalized education plans in France follow a conventional "wait to fail" approach, where children are referred for a personalized education plan only after repeatedly displaying a failure to acquire academic skills. This approach fails to address the subset of children performing below their capacity due to subsyndromic emotional, behavioral, or learning problems.

Additionally, while the efficacy of universal psychosocial skills programs has been established, the adoption rate in France remains low. This could be caused by the lack of practical implementation details provided by the government to school administrators.

The Canopé platform shows promise in addressing the above challenges by offering resources for assessments of students' needs and providing intervention recommendations for teachers. However, as it is targeted at teachers directly, and not integrated into the educational process, it is not systematically applied. The full assessment is lengthy, and the teachers are encouraged to administer only parts of the assessment that they consider relevant to the evaluated student. Teachers have reported insufficient information on how to identify mental health symptoms in children, potentially leading to the oversight of symptoms they are not aware of.

To address the above challenges, we propose a framework that would help teachers (and in the future, parents) use their unique insight into students' wellbeing to address needs of individual students in a systematic fashion.

The goal of the framework is assessment of school students' well-being and learning needs as a basis for providing appropriate adaptive strategies for teachers and parents.

As the first step of the framework, I collaborated with teachers, researchers, and clinicians to adapt an assessment tool (FACETS, see details below) that would offer a comprehensive view of a child's functioning, encompassing mental health, learning abilities, behavior, cognition, and emotion. It offers a common language for communication and collaboration between educators, parents, clinical experts, and other stakeholders. This questionnaire is intended to be completed by teachers seeking to better understand a students' profile (strengths and needs) and support them accordingly. The purpose of the FACETS is to screen for mental health, learning, behavioral and other problems so that teachers can make timely, early, or preventive interventions as well as referrals for specialized services. The intention is for teachers to assess *all* the students in their classroom to avoid biases and overlooking children that are in need but present with less obvious or disruptive symptoms. Care was taken to assure feasibility with respect to teachers' capacity to understand and observe all behaviors assessed by FACETS items while minimizing the time needed to an acceptable duration.

FACETS was originally conceived by Dr. Bennett Leventhal, with inputs from teachers and other experts (prof. Richard Delorme, APHP - Robert Debré and prof. Yasser Khazaal, Lausanne University, who besides contributing to refining the FACETS item pool played a key role in the creation of the French version of FACETS). FACETS contains 63 items and is constructed in collaboration between mental health experts and teachers. FACETS serves as a screening questionnaire, empowering teachers to identify strengths and challenges in their students. Most FACETS items are a visual analogue scale, where both ends (left and right response options) are problematic (*i.e.,* weaknesses), and the middle is the typical behavior (or strength) for the expected/typical developmental level of the child. FACETS items are derived from two sources: key symptoms of mental health disorders outlined in DSM-5[16] and ICD-11[17], and areas of concern identified by educators. The behaviors are not specific (or pathognomonic) to a single disorder, and do not represent diagnostic groups. However, these behaviors commonly appear in classrooms and other settings, and are familiar impediments to learning. During the course of FACETS development, teachers helped prioritize items for inclusion largely on the basis of their prevalence and their impact on learning and adaptation. FACETS is split into ten sections. Nine assess different aspects of wellbeing and learning: communication (both verbal and non-verbal), social function, behavior (e.g., impulse control and activity level), emotion, personality, cognition, learning, somatic and sensory function, and daily routines. The last section contains dichotomous items assessing the presence of issues that require special attention. These items were originally assessed with a continuous scale, similar to the other sections, but were converted to a dichotomous format based on teacher feedback. The items were split into the following categories based on their content similarity. Response data obtained during FACETS evaluation studies will be reviewed to potentially update the FACETS structure. An example of the original FACETS items is shown in figures 19 and 20. Figure 20 shows the original design of FACETS, Figure 20 shows the latest version. The current version of FACETS is available in Annex 5.

Figure 19: FACETS item format – original version

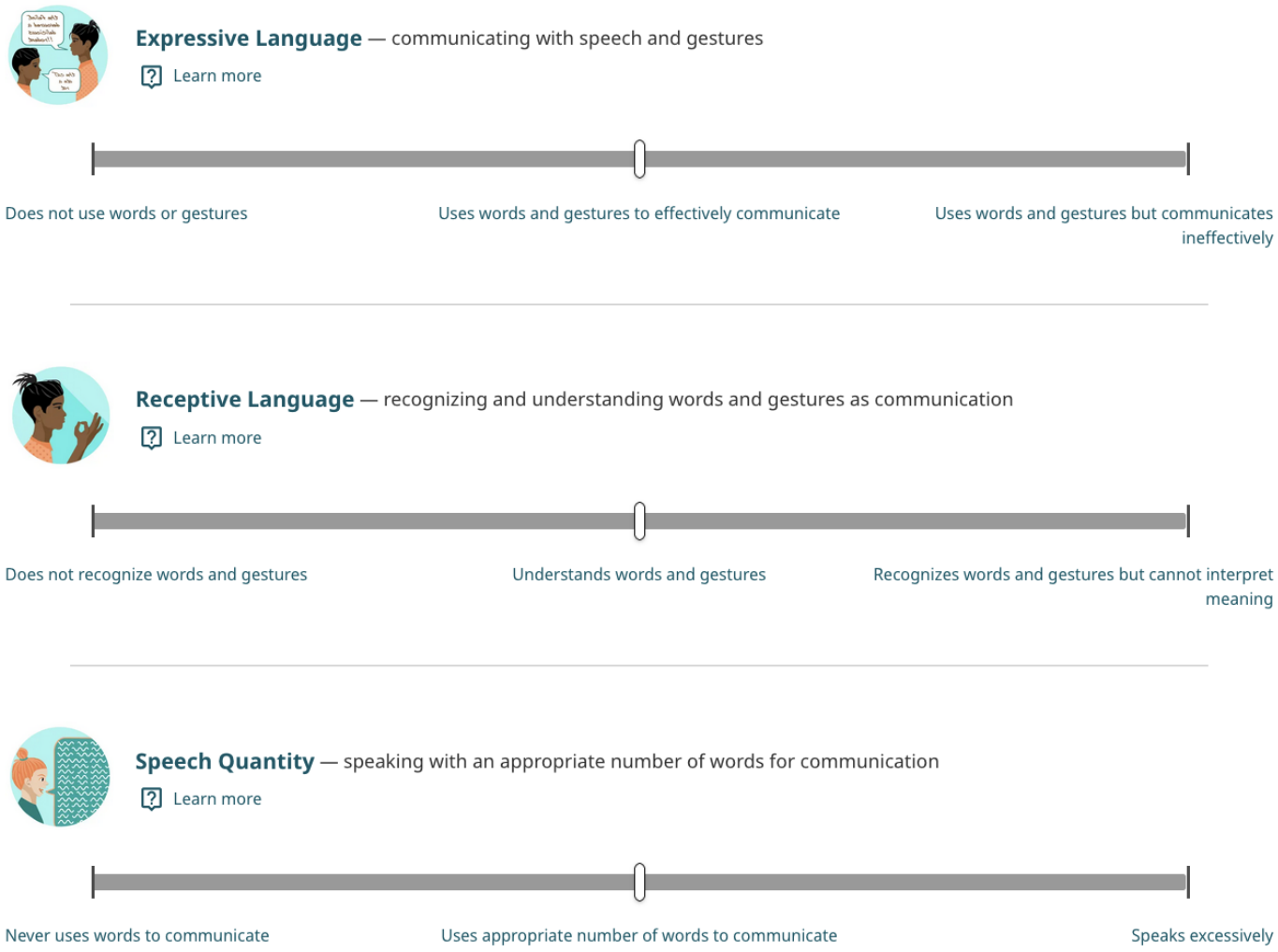

Figure 20: FACETS item format – latest version

The FACETS is built as a web-based tool. It is administered online. Initial production of the FACETS is designed for use through a browser on a computer, tablet, or similar device. A smartphone delivery system is in development. While administration of the FACETS using a “paper-and-pencil” format is possible, the advantage of the online form is that scoring, and then aggregation and visual representation of all available data on an individual child can be readily available upon completion of the FACETS. To facilitate use of the FACETS, tapping on an information icon for any FACETS item will present a brief description of the definition and response options for that item, including examples of the assessed behavior. An example of the descriptions for an item and its response options is shown in Table 9.

Item: Expressive Language (including gestures) — communicating with speech and gestures	Expressive Language includes words and sounds, powerfully augmented by non-verbal communication, including gestures, body posture, facial expression etc. There are many subtle elements that enhance the breadth, depth and nuance of expressive language. These include rate, rhythm, volume, prosody, and tone of speech, as well as vocabulary and word choice. Effective expressive communication requires the flexible, skillful combination of gestures and spoken words. It takes considerable effort to communicate functionally, yet remarkably, even from early life, typically developing individuals are able to express to others their ideas, feelings, wants, and needs. These skills grow in complexity over the course of the lifespan, however, there is considerable variability amongst individuals that depends on biological make-up, cognitive ability, environmental factors, and cultural/social experiences.
Left response option: Does not use words or gestures	Words and gestures are rarely, if ever, used to express preferences or needs. Gestures made with the assistance of others may indicate communicative intent, but are inconsistent. Forced or spontaneous utterances may be present, but are of limited utility due to lack of consistency and use of conventional communication tools such as words or identifiable gestures.
Middle response option: Uses words and gestures to effectively communicate	Speech is characterized by appropriate variability in rate, volume, rhythm, as well as pragmatics, consistent with the context in which the communication is taking place. Uses age-appropriate vocabulary coordinated with visual regard and appropriate gestures (facial expressions, body posture, hand movements, etc.)
Right response option: Uses words and gestures but communicates ineffectively	The lack of coordination of utterances and gestures, along with poor articulation, grammar, and inconsistent rhythm makes expressive language difficult to comprehend and ineffective. Even in the presence of a large vocabulary, expressive language is impaired by the use of words in the wrong context, idiosyncratic variations in vocal pitch (e.g., robotic speech), and rhythm (e.g., stuttering).

Table 9: Example of brief descriptions available for each item and response options.

FACETS allows multiple teachers to assess individual children, ensuring a comprehensive evaluation of their needs. The resulting FACETS report provides a

detailed overview of an individual child's needs, enabling teachers to create tailored interventions for each child.

The FACETS scores of each student are presented on a web dashboard in the form of a "learning profile" of an individual, with responses aggregated across teachers. The report provides the range of scores and median for each item as assessed by the teachers, as well as the individual score of the teacher who is viewing the report. Each individual teacher on their copy of the report will be able to see their score; however, individual scores are not shared with their colleagues. The individual results for all items are displayed on a single page, providing an opportunity for comparing functioning across multiple domains. The goal of the report format is to allow for the rapid identification of distinct individual patterns, for understanding each child's capacities and providing the foundation on which to build necessary interventions.

An excerpt from the FACETS report is shown in Figure 21.

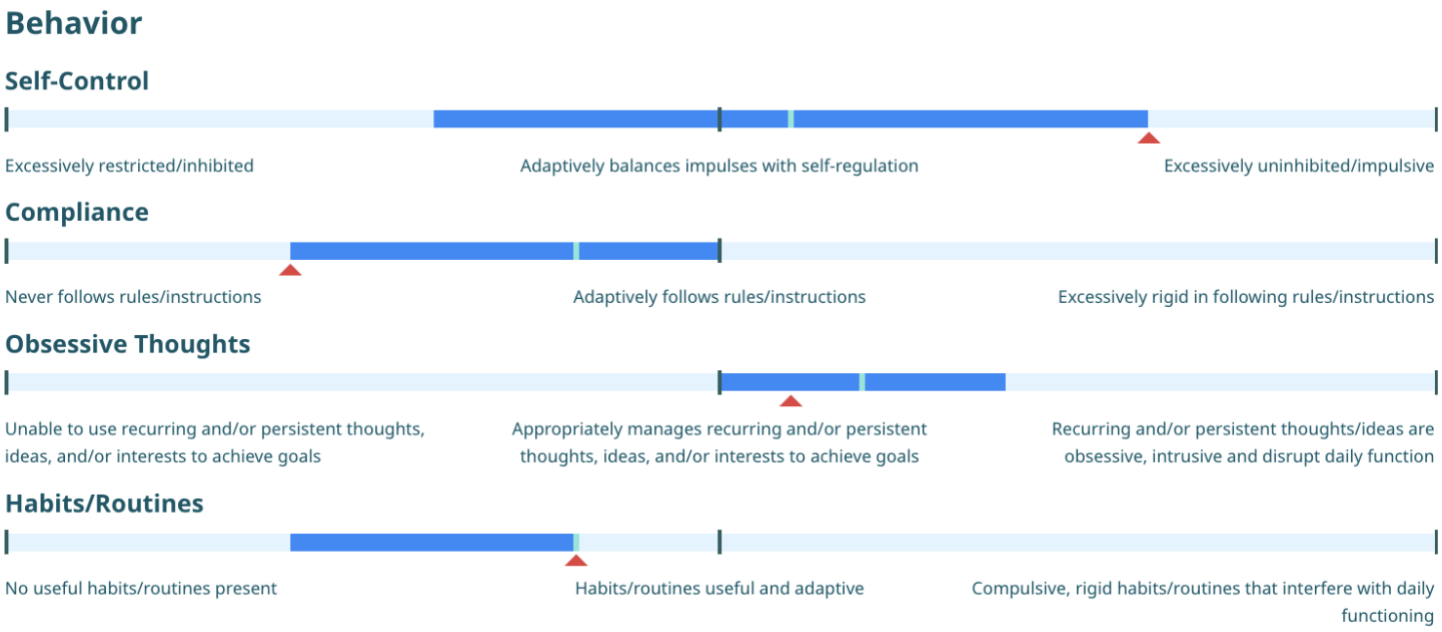

Figure 21: Excerpt from the FACETS report. Range of scores across all raters is shown in dark blue, the cyan line shows the median of scores between the raters, and the response by the rater who is viewing the report is indicated with the red tick mark.

3.2.1. FACETS evaluation

The goal of the current research is to establish acceptance by teachers and psychometric properties of FACETS.

To evaluate the acceptance of FACETS by teachers and obtain preliminary psychometric properties of FACETS (inter-rater and test-retest reliability), a pilot study was conducted in one school.

Acceptance of FACETS by teachers was established with focus-group discussions, and an anonymous survey where teachers were asked to assess each FACETS item. Acceptance of FACETS by clinicians was established with focus-groups discussions, a trial administration of FACETS for one patient, and a structured interview about their experience with the online platform.

Intraclass correlation coefficient (ICC, Fisher, 1992) was used as the main measure of test-retest and inter-rater reliability in the pilot study and will be used in the main reliability study.

There are several versions of the ICC formula. A *one-way* model is used when a different set of randomly selected raters are chosen to rate the participants. A *two-way* model is used when the same set of raters rate the subsets. A *random-effects* model is used when the reliability estimation is planned to be generalized to new raters. A *mixed-effects* model measures reliability for the selected raters and cannot be generalized to other raters with similar characteristics. If the final assessment will use responses from different raters, *multiple-rater ICC* should be used. If measurement from a single rater will be used, then *single-rater ICC* should be used. *Absolute agreement ICC* is used when it is important that the rater assign the same score to participants. *Consistency ICC* is used when it is enough that the two scores are correlated in an additive manner (Koo & Li, 2016).

Koo & Li (2016) defines ICC values between 0.5 and 0.75 corresponding to moderate reliability, between 0.75 and 0.9 to good reliability, and over 9.0 to excellent reliability. Original guidelines are less strict, defining reliability between 0.4 and 0.59 as fair, between 0.60 and 0.74 as good, and over 0.75 as excellent (Cicchetti, 1994).

At the time of the pilot study, the online platform for FACETS administration was not completed, so FACETS was administered using a survey platform that did not allow the integration of brief description of response options. An additional document with the descriptions was made available to the teachers that they could consult when needed.

3.2. Methods

3.2.3. Study setting

The pilot study was conducted at iféa, an innovative school that opened in Paris in 2020. The FACETS project received full support from the iféa board, the teachers, and the parents, following review and obtaining approval from the INSERM Ethical Committee. For this pilot study, 13 teachers participated, each rating up to 6 of their students. 11 parents of 9 students have also participated.

3.2.4. Procedures

To confirm the acceptance of FACETS, I conducted three focus group discussions with the teachers at iféa, three to five teachers in each group. During the focus group sessions, the teachers went through each item of FACETS and were invited to ask questions about the items they did not understand or were not sure about.

As a part of the focus group sessions, a survey was administered, where for each item the teachers chose between three response options, “Accept” if they would like to keep the item, “Reject” if they want to remove the item from FACETS, and “Not sure”.

To examine the reliability of FACETS, FACETS was completed by thirteen teachers, each rating up to 6 randomly selected students. FACETS was filled out twice, from one day to one week apart. One teacher dropped out. Thirty-eight students were rated; Twenty-eight students were rated by more than one rater.

During the pitot study, FACETS was administered using the *SurveyMonkey* (*SurveyMonkey*, n.d.) survey platform.

3.2.5. Consent process

An information form was provided to teachers, parents, and students explaining the details of the study and their role in it. Teachers, parents, and students were informed that if they did not want to participate, they could notify the research team and they would not be included in the study. Teachers, parents, and students were informed that they could withdraw from the study at any point without penalty. Teachers, parents, and students were encouraged to ask any questions they may have prior to participating in interviews. Investigators provided teachers, parents, and students with contact information and were available during the study to answer questions and accept requests to withdraw from the study. Information forms are available in Annexes 6, 7, and 8.

3.2.6. Privacy

Notes of focus-group discussions and results of the anonymous survey administered during the focus group sessions are stored on a secure server. Only the study staff has access to the data. All data will be deleted two years after the most recent publication using this data.

FACETS responses and teachers' email addresses were collected. All information is stored using secure cloud data storage. The data is encrypted and located in the European Union. No unauthorized access to data is given to non-anonymized data. The study research staff does not have access to any identifiable information. FACETS responses were pseudonymized. All students were assigned individual pseudonyms. The key to the code was kept separate in a password-protected file on a private teacher's hard drive or server, accessible only by the teachers. All analysis is conducted only on pseudonymized data.

3.2.7. Data analysis

Focus group data

I calculated how many times each item was mentioned during the focus group discussions, and how many teachers marked the item as "Accepted", "Rejected", or "Not sure" in the focus group survey. I reviewed focus group discussion notes to review teacher's feedback on FACETS as a whole, and identify items that require revision.

Reliability

Parents of several students who were not included in the random selection expressed interest in teachers filling out FACETS for their children. I excluded these students from the analysis to avoid selection bias.

Inter-rater reliability between teachers

Inter-rater reliability was calculated at an item-level. I used only the first entry for each teacher-student combination. I did not use the data from the students rated by fewer than two teachers. Input for the item *Abstract Thinking* is presented in Figure 22.

A	B	C	D	E	F	G	H	J	K	L	M
					18.0	-27.0					
							0.0	33.0			
									0.0	-10.0	
		-1.0									-16.0
2.0										13.0	
					-3.0		7.0				
-3.0		-1.0									11.0
					-44.0	-40.0					
		0.0									0.0
					16.0	-26.0					
9.0				29.0							
	-1.0			2.0							
	-1.0			-8.0							
		0.0	0.0								
3.0				26.0							
		-1.0	1.0								
-1.0				-15.0							
	0.0		0.0								
					-2.0	0.0		1.0			
	0.0			-10.0							
	-1.0		-18.0								
		-1.0	0.0								
					-8.0	-20.0					
										26.0	0.0
					6.0	-15.0					
									-1.0	16.0	0.0
										22.0	0.0
2.0										-18.0	

Figure 22: Input format for calculating the inter-rater reliability between teachers, item "Abstract Thinking". Each column represents a teacher, each row represents a student.

To deal with scale usage heterogeneity (tendencies in how different teachers answer the scale - e.g. all left, all right, all center) I normalized the data by applying this formula to every value in the input dataset: *((value - min_value)/(max_value - min_value)) - 0.5*, where *value* is the original score given by the teacher to the student for the item, *min_value* and *max_value* are the minimum and maximum values that the teacher gave to any student for any item. After normalization, all teachers' value ranges (between all students and items) were between -0.5 and 0.5. I binned the normalized data into three equally sized categories, assigning value "-1" to normalized scores between *-0.501* and *-0.16666667, "0"* to scores between *-0.16666667* and *0.16666667*, and *"1"* to scores between *0.16666667* and *0.5.* One of the items (*Menstruation*) was rated by one of the

teachers on the left extreme for all 6 of their students, and by another teacher for all of their male students. This item was not accepted in the focus group survey by 6 out of 13 teachers. This item was removed from the dataset before normalization.

I used two-way random effects, consistency, multiple-rater intraclass correlation coefficient (ICC) on both original data and normalized binned data. I used the null hypothesis (rho0) of 0.4 as the minimum acceptable reliability. I used the *irrNA* R (Brückl, 2018) library to calculate the ICC, because of the presence of many missing values due to the non-fully crossed design of the study (not all students were rated by every teacher).

To test rater fatigue, I compared the inter-rater reliability on the normalized binned data of the first thirty items and the last thirty items of FACETS using the Mann–Whitney U test.

Using the normalized binned data, I calculated the percent agreement for each item (number of students given the same score by all teachers, divided by the total number of students rated by multiple teachers), average chance agreement, and percent agreement adjusted for chance agreement. To calculate the average chance agreement and p-values of the percent agreement I used a randomization test, calculating the percent agreement on 100 samples with scores reshuffled between students and teachers, preserving the proportion of the scores within each item. To calculate the adjusted percent agreement, I used the formula of Cohen's Kappa: (observed agreement - chance agreement)/(1 - chance agreement). I calculated 95% confidence intervals for the percent agreement and adjusted agreement using bootstrap resampling (200 resamples).

Inter-rater reliability between teachers and parents

I calculated the average score between parent responses to obtain the average parent score for each student on each item, and average scores between teachers to obtain the average teacher score for each student on each item. I used only the first entry for each teacher-student combination. Input for the item *Abstract Thinking* is presented in Figure 23.

Parents Avg	Teachers Avg
10.0	16.5
13.0	2.0
22.5	0.0
10.0	-5.0
-7.0	0.0
-9.0	-9.0
23.0	15.0
28.0	-8.0

Figure 23: Input format for calculating the inter-rater reliability between teachers and parents, item "Abstract Thinking". Each row represents one student.

Each score was divided by 100 to obtain values between -0.5 and 0.5 and replaced with "-1" if the value was between *-0.501* and *-0.16666667, "0"* between *-0.16666667* and *0.16666667*, and *"1"* between *0.16666667* and *0.5.*

I used the *irr R* library (Gamer et al., 2019) to calculate two-way random effects, consistency, multiple-rater ICC on the original data, with null hypothesis rho0=0.2, as the reliability between teachers and parents was expected to be lower than reliability between teachers.

I used the *irr R* library to calculate unweighted Cohen's kappa on the binned data. We used the *psych R* (Revelle, 2023) library to calculate confidence intervals for the kappa values.

As an alternative measure of inter-rater reliability, I calculated the Spearman's rank correlation coefficient between the average parent score and the average teacher score on both original and binned data.

Test-retest reliability

Some of the teachers filled out FACETS more than two times for some of their students. One teacher filled out the FACETS on a non-randomized list of students. In this case, I skipped the erroneous entries, and kept only the first and the last entry for each student.

I skipped all entries where a teacher only filled out FACETS once for a particular student, and teachers who rated <4 students.

Input for the item *Abstract Thinking* and teacher A is presented in Figure 24.

Score 2	Score 1
4	2
4	-3
26	9
5	3
3	-1
2	2

Figure 24: Input format for test-retest reliability calculation, item "Abstract Thinking", teacher A

For data normalization and binning, I used the same procedure as for inter-rater reliability, except I used the data from both FACETS administrations to calculate the teacher score ranges.

I calculated test-retest reliability for each item/teacher combination using single, 2-way mixed-effects, absolute agreement ICC, with rho0 = 0.4 on the original (non-normalized) data, using the *irr R* library. I calculated the average of the ICC values per item.

I used the *irr R* library to calculate unweighted Cohen's kappa on the normalized binned data. We calculated the average kappa values per item. I used the *psych R* library to calculate confidence intervals for the kappa values.

I used the Mann-Whitney U test to check if the kappa values for teachers more fluent in English is higher than the kappa values between all teachers (to test if English fluency affected reliability), if the kappa values for teachers who spend less time with the students is lower than the kappa values between all teachers (to test if the amount of time spent with the student affected reliability), if the kappa values over all teachers and items for entries with the inter-administration time in minutes under the mean (~5.4 days) is higher than the kappa values for entries with the inter-administration time in minutes over the mean (to test if the time between the two administrations affected reliability), and if the

kappa values for the first thirty items is higher than for the last thirty items (to test potential rater fatigue).

As an alternative measure of test-retest reliability, I calculated the Spearman's rank correlation coefficient (r_s) between the two administrations on both original and binned data. I calculated the average r_s values per item.

Administration time analysis

I used the Mann-Whitney U test to check if the time it took teachers to fill out FACETS (administration time) was longer for the first administration of FACETS for a particular student than the second administration. Several teachers mentioned taking breaks while completing FACETS, occasionally continuing the task on the following day. These breaks were included in the platform's administration time. To address this issue, we excluded administration periods exceeding one hour to minimize the impact of extended break times on the overall completion process.

I calculated the mean administration time in minutes for each FACETS administration for any student (twelve total administrations for most teachers, six for the first administration and six for the second administration). I calculated the Pearson correlation coefficient between FACETS administration time and the administration count among the first twelve administrations.

I used the normalized binned data from the first FACETS administration (used to calculate inter-rater reliability between teachers), only containing students rated by two teachers to calculate the Pearson correlation coefficient between the average administration time for the student and the sum of absolute scores between all items and teachers (i.e. number of "extreme" values) of the student. I did not use entries with administration time longer than one hour for this analysis.

I fit a negative exponential function (*a * np.exp(-b * x) + c*) to administration times of the first administration for each student to find the asymptote administration time.

Cluster analysis

To explore relationships between items, hierarchical clustering was performed on unaggregated data, using the *stats* package in R (*hclust* and *cutree functions)*.

3.3. Results

3.3.1. Acceptance of FACETS by teachers

All questions were accepted by the majority of the teachers. No questions were rejected by more than one teacher. Three questions were rejected by one teacher: *Social Communication*, *Persistent Thought* and *Sleep*. Some questions were marked as “Not sure” by more than three teachers, among them *Substance Use* (4), *Future Outlook* (3), *Menses* (6), *Morning Routines* (3), *After-School Routines* (3), *Going Out Routines* (3), three of those being from the same “Daily Routines” category. *Menses* was marked as “Not sure” by 6 teachers, more than any other item. With an exception of *Menses*, the items that the teachers asked clarifications for during the focus group discussions did not overlap with the questions that were marked as “Not sure”.

3.3.2. Reliability of FACETS

Although invisible to the teachers, all the items were rated on a scale between -50 and 50. All item means are between -10 and 10, most skewed items being *Screen Time* (8.4), *Sexual Behavior* (7.8), and *Menstruation* (-7.5). Four items had a standard deviation below 4: *Substance Use* (3.98), *Toileting* (3.3), *After-school routines* (3.3), and *Going out routines* (2.5). All four of these items are among those that were marked as “Rejected” or “Not sure” on the pre-focus group survey by multiple teachers.

Visual examination of the teacher responses showed that there were very few items that are rated on opposite extremes by two teachers, there is high agreement on the items rated close to 0, the teachers often agreed on the direction but not the magnitude of a behavior, and that the magnitude difference is often systematic (some teachers always give more “extreme” scores than others). An example comparison of scores between two teachers for one student on a subset of FACETS items is presented in Figure 25 (left). To account for the difference in magnitude of scores between teachers, I normalized the data per teacher, so that the scores of each teacher fall within the same range, and then binned the resulting score in three equally sized categories, representing the left extreme, typical behavior, and the right extreme (Figure 25, right).

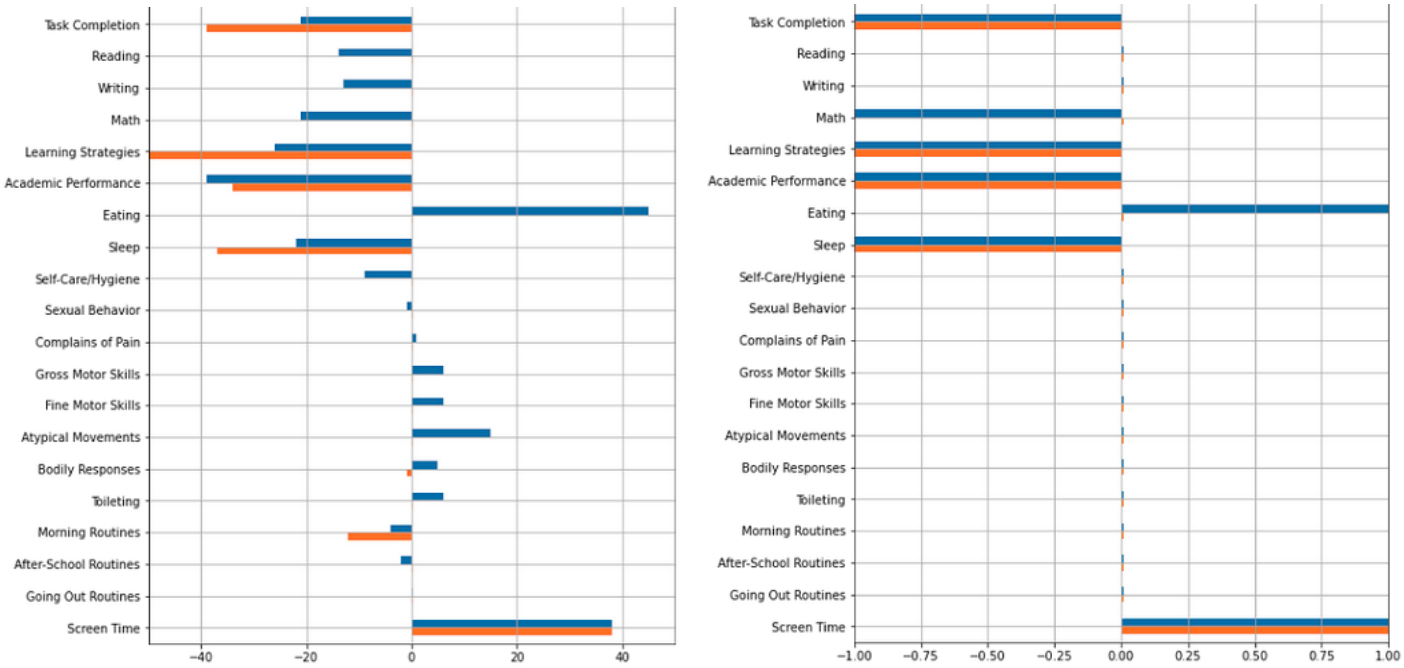

Figure 25: Original (left) and normalized and binned (right) scores of two teachers for one student on a subset of FACETS items. Invisible bar indicates that the teacher scored the item as 0.

Inter-rater reliability between teachers

The average ICC for all items on the original data was 0.19 on both the original and the normalized binned data. On the binned data twenty-one items have fair inter-rater reliability (ICC>0.4), with items *Academic Performance*, *Social Communication*, *Aggression*, *Sleep*, and *Speech Quantity* having lower CI bound >0.24.

No significant difference between inter-rater reliability on the first thirty items of FACETS and the last thirty items was found, suggesting the absence of rater fatigue.

Adjusted percent agreement was interpreted using the same bands as Cohen's Kappa since the same formula was used for its calculation. Fourteen items have adjusted agreement over 0.21 (fair), seventeen items having lower CI bound >0.1.

Inter-rater reliability between teachers and parents

Nineteen items have an ICC over 0.4 (original data), *Integrity* having lower CI bound 0.12. Seven items have kappa over 0.21.

Test-retest reliability

The ICC value for the original data varied between teachers and items (e.g. between 0.97 and -0.25 for *Abstract Thinking*, between 0.9 and -0.79 for teacher A). The average value over all teachers and items was 0.35. Twenty-seven items have average test-retest reliability over 0.4.

Cohen's Kappa (κ) could not be calculated for five items due to absence of response variance. Forty-three items have fair average test-retest reliability (κ>0.21)[2], nine items having the average lower CI bound of the test-retest reliability of κ>0.21.

Average κ was not significantly higher for teachers more fluent in English. Average κ was significantly lower among teachers who spend less time with the students (p-value=0.0007).

κ values for the entries with the inter-administration range below the mean were significantly higher than the values for the entries with the inter-administration range over the mean (p-value=0.02).

κ values for the first thirty items were not found to be higher than the values for the last thirty items.

Administration time analysis

Administration time reduced significantly between the first and the second administration of FACETS for the same student, from eleven to five minutes (after removing administration times longer than one hour, p=0.00005). Figure 26 shows the progression of administration times for all teachers. Only the first twelve administrations are shown, as most teachers filled out FACETS twelve times (twice for each of their six students).

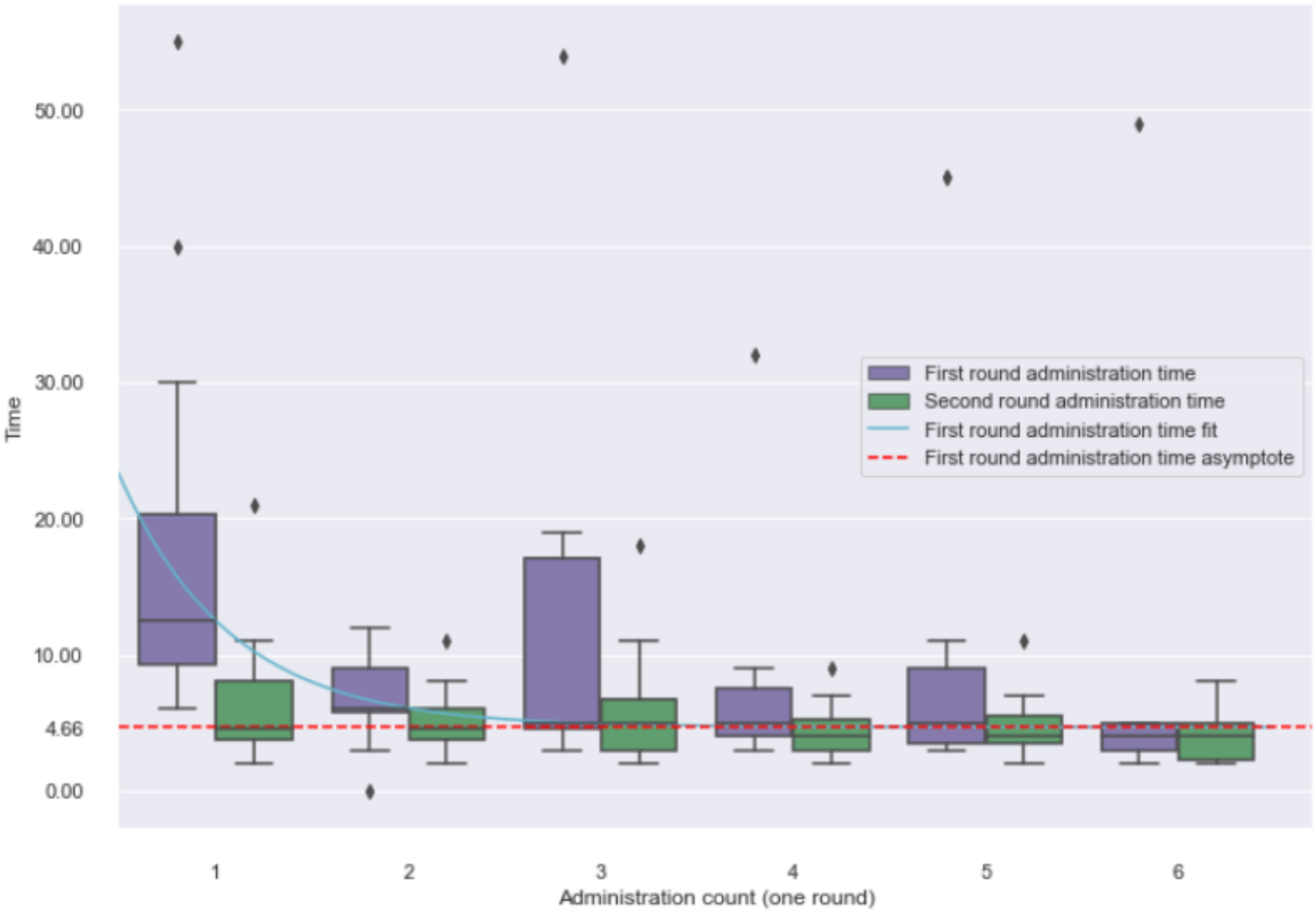

Figure 26: The graph shows the progression of administration times. The box represents the quartiles, with the line inside representing the median. The whiskers show the range of the data, except for any points that are further than 1.5 times the interquartile range from the edges of the box, which are plotted as separate dots. The purple blocks represent the first FACETS administration for the same student, while the green block represents the second administration for the same student. A logarithmic function is fitted to the first administration of FACETS, shown in blue. The red dotted line represents the asymptote administration time.

The Pearson correlation coefficient between the mean administration time and administration count (after removing administration times longer than one hour) is -0.76, p-value=0.004. The Pearson correlation coefficient between the average administration time for the student and their number of “extreme” scores was 0.37, p-value=0.055. The asymptote administration time was found to be 4.66 minutes.

Cluster analysis

Figure 27 presents the hierarchical clustering dendrogram on the teacher responses.

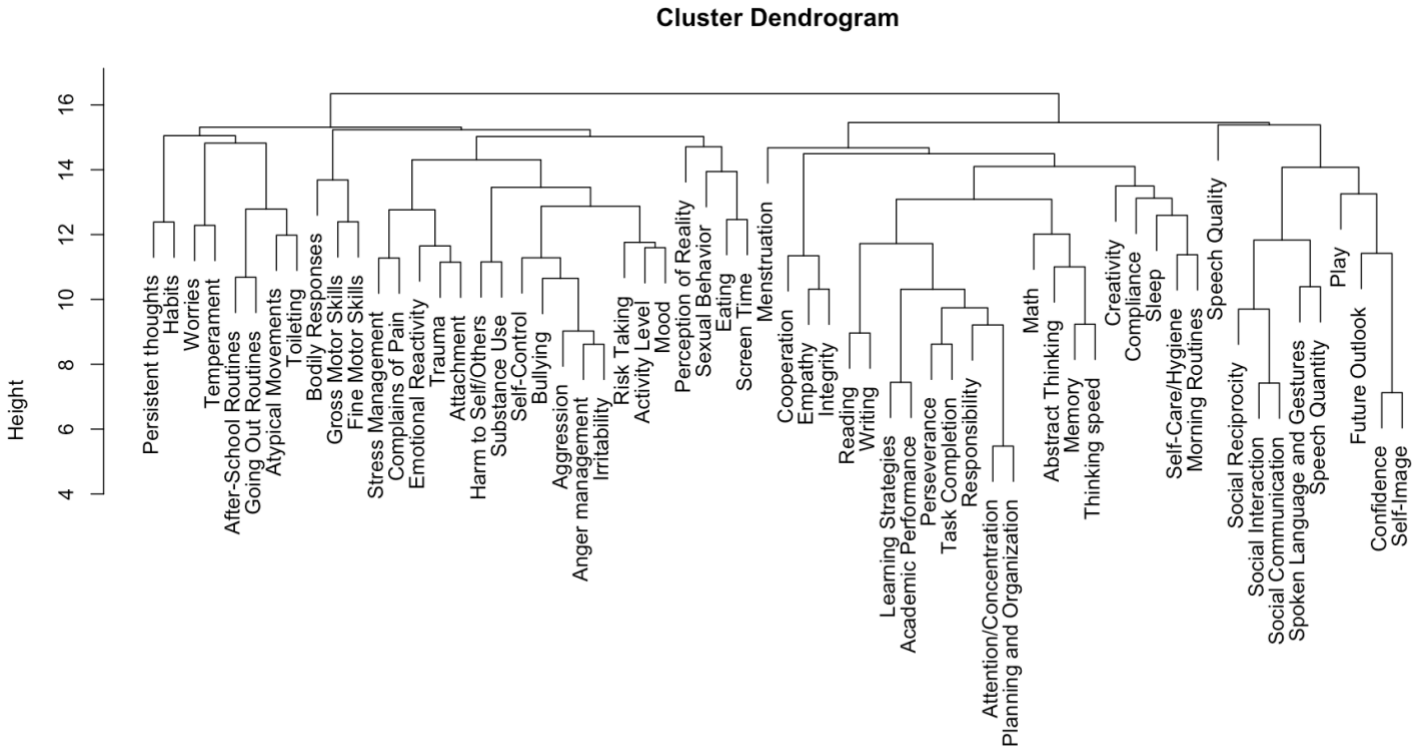

Figure 27: Hierarchical clustering dendrogram of the teacher responses.

3.4. Discussion

The focus groups sessions conducted as a part of the pilot study indicated overall acceptance of FACETS and helped highlight items that needed to be improved for the current study. Statistical analysis of FACETS administration data showed low variance in scores on items about which teachers expressed a concern during focus group sessions, possibly confirming the teachers' concern of not being able to assess the behaviors.

Inter-rater and test-retest reliability was low to average. The sample size was insufficient to obtain meaningful confidence intervals for most items. There was no significant change in inter-rater and test-retest reliability between the first and second part of FACETS, indicating the absence of rater fatigue and suggesting that there is no need to reduce the number of FACETS items. The time the teacher reported spending with the student had a significant effect on the test-retest reliability, however teachers' English proficiency had no significant effect. Test-retest reliability values reduced as the time between administrations increased, which can be explained either by the change in the measured

variables or potential learning effects. Fixed time between the two administrations is warranted for future studies.

Administration times significantly reduced with each consecutive FACETS administration. Students with more "extreme" scores take more time to assess. Asymptote analysis revealed that the administration time decreased to less than five minutes within the first six FACETS administrations.

Hierarchical clustering of the unaggregated teacher data on the limited sample available during the pilot study showed expected clusters (e.g. Self Confidence item grouped together with Self-Image, both grouped with Future Outlook), suggesting that the teachers understood the items.

In conclusion, while the pilot study showed an overall acceptance of FACETS, observations of reliability and teacher feedback prompted a revision of the instrument. Key improvements include the development of a French version that would allow administration in a wider range of schools in France, rewording of items and response options based on teacher feedback, converting some continuous items to a dichotomous (Yes/No) version, and the introduction of a middle response option describing the typical/expected behavior. Additionally, a new web platform that integrates brief descriptions of the items and response options directly into the administration process is expected to improve item reliability. A larger sample size is required to obtain reliability estimates. A new version of the English version of FACETS administered with the online platform is presented in Figure 28, with the brief description of the middle response option of the Self-Control item activated.

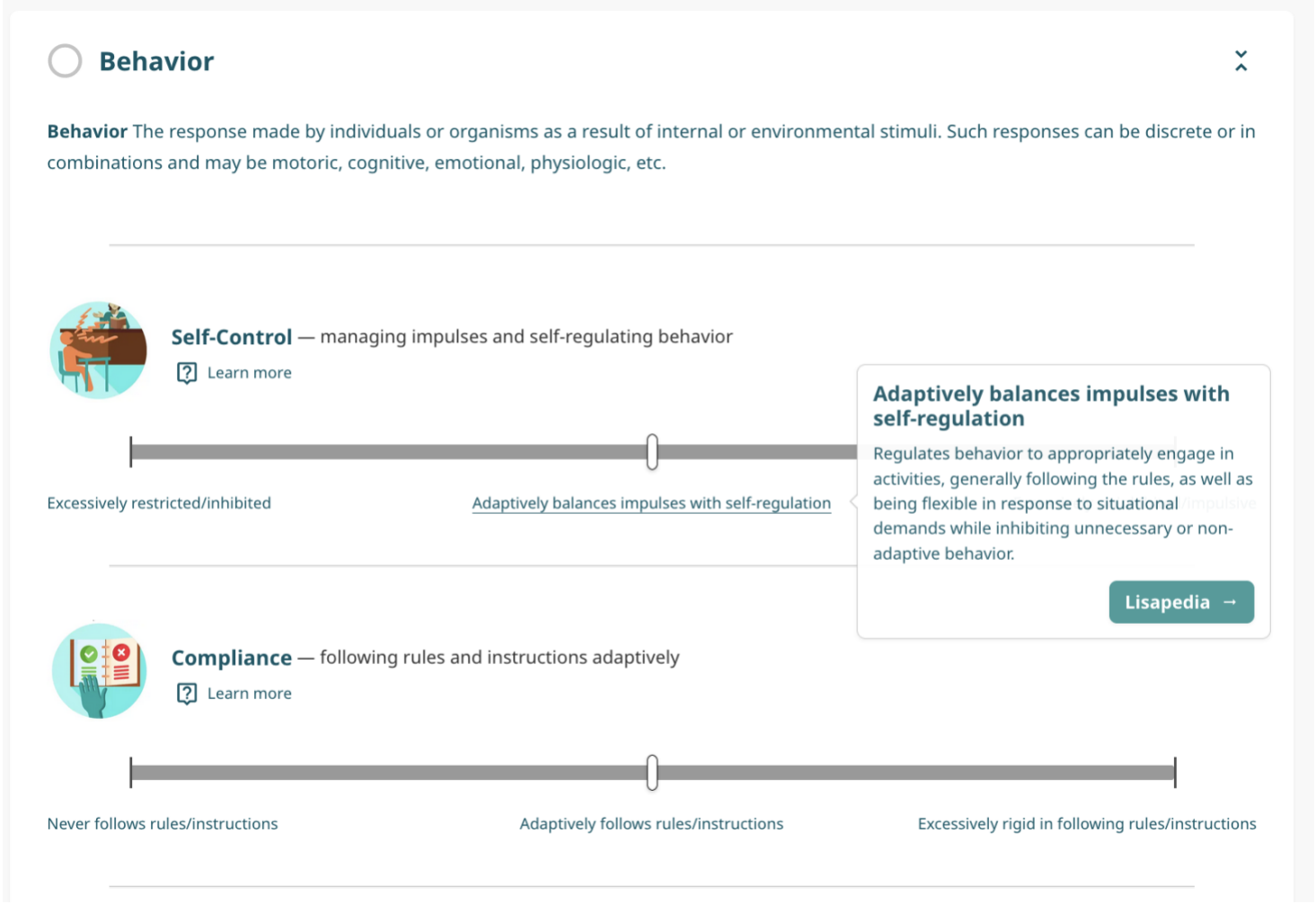

Figure 28: Excerpt from the new version of FACETS.

3.5. Ongoing work

3.5.1. Acceptance, reliability, and factor structure study

A bigger study with a larger sample size is currently being conducted to obtain inter-rater and test-retest reliability values for the new version of FACETS,. At the time of writing the first phase of the study was completed (establishing acceptance of FACETS by the teachers). A second phase is planned for the end of November 2023, where I will establish the reliability and factor structure of FACETS.

3.5.1.1. Sample size estimation

For inter-rater reliability between teachers, the acceptable ICC value was estimated to be between 0.4 and 0.7, since the raters are not equal – they see the child in different

contexts (teachers teach different classes). According to Bujang (2017), with these conditions, for a power of 80% and a statistical significance of 0.05, if each teacher filled out the FACETS for five students, twenty-nine students rated by the same teachers are required.

For the inter-rater reliability between the teachers and the parents (two observations per student: one parent and the average of scores from the teachers for each student), the acceptable reliability was estimated to be between 0.3 and 0.6. With these conditions, for a power of 80% and a statistical significance of 0.05, forty-four students are required.

The inter-rater reliability between clinicians is expected to be between 0.6 and 0.8. With these conditions, for a power of 80% and a statistical significance of 0.05, a sample of at least thirty-nine students is required.

After reliability of FACETS is confirmed, factor structure will be examined using exploratory factor analysis, which requires a minimum of fifty students (de Winter et al., 2009).

3.5.1.2. Methods

Study setting

The study is conducted at two schools of the iféa network.

Procedures

To confirm the acceptance of FACETS, I conducted two focus group sessions with the teachers, one in each school, following the protocol from the pilot study. The FACETS was updated according to feedback from the teachers.

During the next phase of the study, FACETS will be completed by 11 teachers in two schools, each rating 60 students on two occasions, from one day to one week apart.

Consent process and privacy

The consent process and privacy considerations are identical to the pilot study.

Data analysis

The focus group sessions followed the protocol from the pilot study.

Each item will be z-score normalized and binned based on the z-scores. I will perform the reliability analyses on both raw data and normalized binned data.

Inter-rater reliability will be calculated using two-way random effects, agreement, multiple-rater the intraclass correlation coefficient (ICC) with null hypothesis rho0=0.4 (minimum acceptable reliability), using the *irr R* library, and the Spearman's rank correlation coefficient (r_s) on both raw and normalized binned data.

I will calculate the average score between parent responses to obtain the average parent score for each student on each item, and average scores between teachers to obtain the average teacher score for each student on each item. I will use only the first entry for each teacher/student combination.

I will use two-way random effects, agreement, multiple-rater ICC with null hypothesis rho0=0.3 and Spearman's rank correlation coefficient (r_s) on both raw and normalized binned data.

I will calculate test-retest reliability for each item/teacher combination using two-way random effects, agreement, multiple-rater ICC and r_s on both raw and normalized binned data.

For the teacher data, I will aggregate the data for each teacher using the median score of each of the respondents' scores, and check if the data is suited for EFA (exploratory factor analysis) using Kaiser-Meyer-Olkin (KMO) and the Bartlett's Test of Sphericity tests. I will perform factor analysis on the parent and teacher data separately. I will use principal axis factor (due to non-normality of the data) and oblimin rotation (to avoid a priori assumption that the items do not correlate with each other). A scree plot will be used to identify the number of factors.

Hierarchical clustering will be performed separately on the aggregated parents' and teachers' responses using *stats* package in R (*hclust* and *cutree functions)*.

The administration time analysis will be repeated following the pilot study protocol.

3.5.1.3. Results

All items were accepted by the majority of the teachers. Six items were rejected by two teachers. Two items were rated as “Not sure” by more than two teachers. Items were revised according to the teacher feedback (e.g. the wording of items made applicable to a wider range of developmental levels, response options for items in the learning category have been changed to include behaviors often observed by teachers but not captured in the current wording of the items).

3.5.1.4. Conclusions

Focus group discussions and the results of the focus group survey indicated overall acceptance of FACETS. Minor adjustments have been made to the scale based on teachers’ feedback.

3.5.2. Acceptance, reliability, and validity study in clinical settings

One of the goals of FACETS is to act as a universal screening instrument, intended for equitable identification of symptoms that warrant referral to mental health services outside school. To examine whether FACETS items are valid indicators of mental health symptoms, a validity study is underway that will examine convergent and discriminant validity of FACETS items with assessments routinely administered in child psychiatry hospitals in France.

The current study is taking place at a child psychiatry department at the Robert-Debré hospital in Paris. The department comprises five clinics dedicated to general child psychiatry, learning disorders, eating disorders, autism spectrum disorder, and attention deficit hyperactivity disorder. A number of screening and diagnostic assessments are routinely administered at each clinic, including parent- and self-report scales and cognitive task batteries (for example, Behavior Rating Inventory of Executive Function and Hamilton Anxiety Rating Scale). Each FACETS item has been mapped to a subscale from one of the assessments (e.g. Attention and Activity level items to the ADHD Rating Scale-IV assessment).

In addition to examining convergent and discriminant validity, I will additionally confirm inter-rater reliability by comparing FACETS scores between senior and junior clinicians.

The first phase (clinician acceptance) of this study is completed, the second phase will start at the end of November 2023.

3.5.2.1. Sample size estimation

I will examine convergent and discriminant validity of FACETS items by calculating correlation coefficients between FACETS items and corresponding subscales from standard assessments. I estimate acceptable correlation with standard assessments to be 0.55. According to Bonett & Wright (2000), for a confidence level of 0.95 and a confidence interval width of 0.4, the required sample size is fifty-seven patients. To confirm the discriminant validity of FACETS items, acceptable correlation with standard assessments is estimated to be close to 0. According to Bonett & Wright (2000), for a confidence level of 0.95 and a confidence interval width of 0.55, the required sample size is fifty-two patients. To account for possible participant drop-out, of 25%, clinicians will conduct FACETS on seventy patients.

3.5.2.2. Methods

Study setting

The study is being conducted at the child psychiatry unit at the Robert-Debré hospital in Paris.

Procedures

Acceptance of FACETS by the clinicians

Together with my intern, Pernille Brams, we convened a focus group with three clinicians, where they were introduced to the project, filled out the FACETS for one imaginary patient using the online platform, and reviewed the resulting profiles on the dashboard. The administration process was observed by the coordinator, taking note of the understanding of FACETS usage by the clinicians, time taken to complete FACETS, the ease of use of the online platform, and any bugs identified during the session. After the trial FACETS administrations, the clinicians were administered a structured interview about their experience with FACETS and the online platform.

Reliability and validity of FACETS

FACETS will be completed by clinicians at the Robert-Debré Hospital. FACETS will be administered by clinicians in five departments (ADHD, Learning, Eating disorders, General child psychiatry, ASD). Each junior clinician at each department will complete FACETS for two patients per week, four patients per clinic, for fifty weeks. The senior clinician of each department will complete FACETS for two patients assessed by one of the junior clinicians per week.

Consent process and privacy

The consent process and privacy considerations are identical to the pilot study. There will be no changes to how the routine hospital assessments are administered and stored.

Data analysis

Each item will be z-score normalized and binned based on the z-scores. I will perform the reliability analyses on both raw data and normalized binned data. Inter-rater reliability between senior and junior clinicians will be calculated using the intraclass correlation coefficient (ICC) and the Spearman's rank correlation coefficient (r_s) on both raw and normalized binned data.

I will calculate Pearson's and Spearman's rank correlation coefficients between FACETS items and corresponding subscales from standard hospital assessments.

3.5.2.3. Results

Establishing acceptance of FACETS by clinicians

Clinicians generally found the online platform to be intuitive and easy to use. Some minor bugs were noted. Feedback on the FACETS indicated that it was reasonably concise. The items and response options were found to be effective in capturing a broad spectrum of relevant aspects of children's behavior. Two items were found to be challenging to evaluate (*Emotionally and Physically Harms Others* and *Integrity and Honesty*).

3.5.2.4. Conclusions

The focus group discussions indicated overall acceptance of FACETS by clinicians. Minor changes have been made to the online platform and the FACETS items to address clinicians' feedback.

3.6. Discussion

Focus group discussions among both teachers and clinicians expressed overall acceptance of FACETS, leading to minor adjustments based on feedback. Administration times decreased with consecutive assessments, with an asymptote of less than five minutes reached within six administrations. Clinicians at the Robert-Debré hospital indicated that the length of FACETS was acceptable. Studies assessing reliability, validity, and factor structure of FACETS will begin in November 2023. In case of insufficient psychometric support FACETS will be revised further in collaboration with teachers and mental health experts.

While the studies showed good reception of FACETS among both teachers and clinicians, it is important to acknowledge certain limitations. The teachers and clinicians involved in the pilot study may not fully represent the diverse perspectives and practices found in various institutions. The acceptance of FACETS was assessed within a specific cultural and educational setting, and its applicability in different cultural contexts remains to be evaluated. Furthermore, the reliance on focus group discussions introduces the possibility of social desirability bias, where participants may provide responses that align with perceived expectations. Future studies will extend the findings in diverse educational settings and explore different evaluation methods.

During informal discussions with teachers, they reported that FACETS was helpful in enhancing their understanding of their students' strengths and weaknesses by systematizing their observations of students' behavior and providing a shared vocabulary for discussions among teachers where appropriate actions are often discussed. FACETS also seemed to alleviate teachers' anxiety about challenges in their classrooms by offering explanations for concerning behaviors. The validity of these observations will be examined in the upcoming studies.

3.6.1 LISA framework

FACETS has become a part of a larger LISA framework (Learning, Integration, Support, Awareness), dedicated to deploying evidence-based learning strategies for educators to support the healthy development of children in schools, co-initiated and directed by Dr. Elie Rotenberg. The longer-term goal of the LISA framework is to build seamless transitions from assessment and creation of a learner profile to the identification and implementation of classroom strategies.

Besides the FACETS assessment, the LISA framework will include a dashboard and a report center. The dashboard will allow analysis of patterns of FACETS responses. Similar to the Canopé online tool, the report center will show insights into a child's strengths and needs, as well as evidence-based intervention strategies (Figure 29).

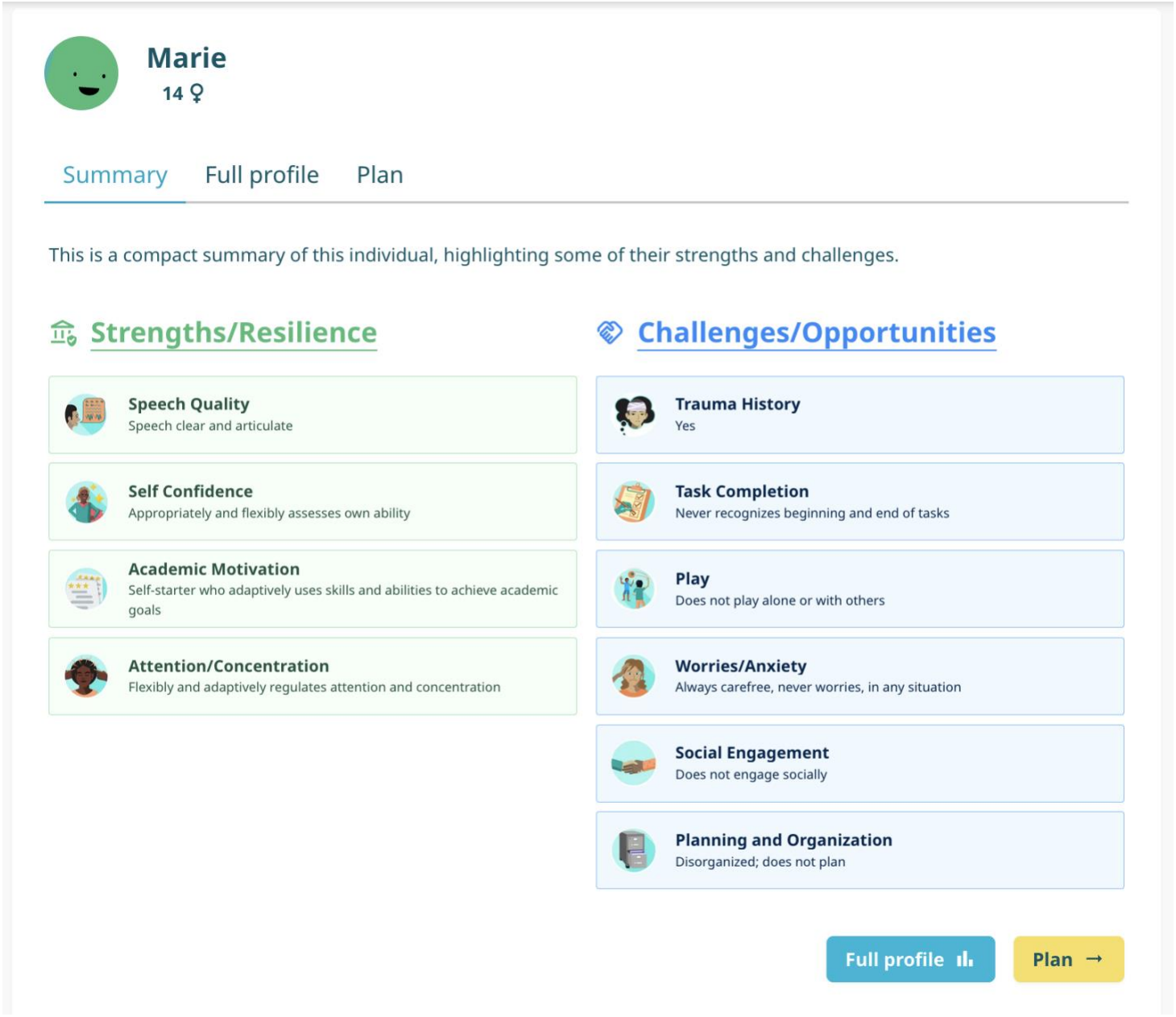

Figure 29: Example of a report center summarizing strengths and needs of a student.

The LISA framework will also include LISA-DB (LISA Mental Health and Learning Psychoeducational Resources Database) – a collection of practical guides for parents and teachers. The LISA-DB compiles practical guides developed and vetted by experts, offering parents and teachers a structured approach to addressing behavioral concerns identified through FACETS.

The Guides are designed to provide a framework for approaching problems identified by the FACETS. Each guide has a similar structure beginning with the description of the nature of the behavioral concern. These descriptions include examples of not only typical behavior but also behaviors along the continuum captured by the FACETS. With these descriptors in mind, users are then provided with possible strategies for addressing the behavioral concern. In the case of undesirable behaviors, suggestions are provided for reducing their frequency, intensity, and/or level of interference with functioning. Similarly, for desirable behaviors, strategies are suggested for increasing frequency and appropriateness of function in these areas. The guides are currently being developed in collaboration with multiple clinicians and mental health researchers.

To integrate proposed interventions into the educational process, LISA is developing pre-built educational plans and templates and planning tools that allow for the creation of individualized and classroom-level educational plans, and continuous monitoring of their implementation and effectiveness.

Structured training programs for teachers are planned, to equip educators with the knowledge and skills necessary for effective support of individual students. While the primary training focus is for teachers, LISA provides training for parents as they play their critical role in guiding and supporting the child’s education, along with training children to be active participants in cooperative and peer-to-peer education.

LISA expands its network by involving additional psychologists and psychiatrists, ensuring appropriate expertise is available to support the personalization of the educational process.

Integrated into each school's instructional team and calendar, LISA will provide systematic assessments for all students to identify students' strengths and needs in different dimensions, and to take into account the wide variations among developing young people, including those with special needs/students who struggle but fall below the thresholds for a disorder diagnosis. This will allow timely identification of children in need of PPRE, PAP, or PPS support. Additionally, LISA training workshops will help teachers identify problems requiring PPRE earlier, before they significantly interfere with educational achievement.

LISA will ensure a seamless, fully digitized transition from identification to implementation of learning strategies. The LISA report will integrate all the elements suggested/required by the PPRE document, and the GEVA-sco document of the PAP. It will go beyond these elements to also integrate the behavioral, cognitive and social profile of the student evaluated by FACETS. It will highlight areas for the development, management, and reinforcement of skills, notably executive functions (planning and organization, etc.); emotional management (worry, frustration, etc.); and social functions (cooperation, communication, etc.). LISA will provide planning templates (with pre-filled examples) for individual and classroom scenarios on how strategies and interventions could be designed and implemented.

In July, 2023, Pap Ndiaye, Minister of National Education and Youth, and Bruno Bonnell, Secretary General for Investment, announced LISA as the winner of the "Innovation in school form" Call for Expressions of Interest (l'Appel à Manifestation d'Intérêt, «Innovation dans la forme scolaire»), with a grant of 2 million euros for the duration of 5 years. The LISA framework will expand to over 230 pilot establishments in the Ile-de-France academic region, benefitting more than 80,000 students, and be gradually integrated within the national educational system. This expansion will enable larger studies in more diverse samples, with the goal of ensuring the utility of the LISA framework and further confirming the psychometric properties of FACETS.

4. Discussion and perspectives

4.1 Discussion

The presented thesis addresses the challenge of the detection and support of mental health and learning difficulties in children and adolescents. The two distinct approaches, FACETS questionnaire, as a part of the LISA framework, and the data-driven method using the Healthy Brain Network (HBN) dataset, offer potential solutions to the existing gaps in current screening practices.

The application of feature selection techniques to the HBN dataset shows the potential of leveraging existing data to construct efficient screening tools. The machine learning models trained on this dataset show improved performance over existing assessments, providing an alternative approach for the development of new assessments that requires less novel data collection. The flexibility of this approach offers adaptability across diverse populations and clinical applications.

Several studies applying feature selection techniques have been published during my work on this project. Schultebraucks et al. (2023) predicted future post-traumatic stress disorder symptoms among emergency patients by reducing a 27-item screener to 5 items using Recursive Feature Elimination. Tutun et al. (2023) used a filter method (Networked Pattern Recognition) to reduce the number of items in a multi-disorder screener. Glavin et al. (2023) used exhaustive feature selection to identify two items from a 9-item screener for depression screening in primary care.

The application of feature selection techniques to the HBN dataset provides a further contribution to the field by using item-level responses from a larger number of assessments compared to other similar studies. This extensive item set allows for a comprehensive exploration of the assessment space, enabling the identification of relevant items across assessments targeting various disorders and different rater types. Considering the substantial number of input variables, computational complexity was a crucial consideration. This research extends prior work by validating the application of previously identified techniques to mental health assessment datasets characterized by a high number of variables. This empirical validation supports the robustness of the applied methodology when applied to datasets with a substantial variable count, as encountered in recent deep phenotyping data collection projects, such as the Nathan

Kline Institute-Rockland Sample (Nooner et al., 2012) and the Adolescent Brain Cognitive Development study (Bjork et al., 2017).

During the work on the HBN analysis, I discovered recently that another team (Senseable Intelligence Group at the Massachusetts Institute of Technology) was conducting similar research using the HBN dataset and machine learning models (unpublished results shared by S. Ghosh) to predict ADHD diagnosis using assessment responses resulted in similar classification performance. This convergence of results across independent research projects confirms the efficacy of employed methodologies. My work presented in this thesis provides a general approach that was applied for prediction of multiple diagnoses.

The collaborative development of FACETS, the broad-ranging screening scale designed to identify the needs of school children, underscores the importance of engaging both teachers and clinicians in the creation of screening tools intended for use in the educational setting. Addressing the behaviors encountered by teachers in the classrooms, using vocabulary familiar to teachers, and accommodating time constraints faced by teachers were all tackled through close collaboration with teachers in iterative cycles, working in conjunction with clinical experts. While the pilot study indicated an overall acceptance of FACETS by teachers, the initial reliability estimates were suboptimal.

To enhance teachers' comprehension of FACETS items, both items and response options were reworded based on feedback from the teachers, and introduced a middle response option describing the typical or expected behavior. Additionally, a French version of FACETS was developed, and the online platform was finalized, which integrates brief descriptions of items and response options into the administration process, aiming to further improve teachers' understanding of the items.

FACETS, as a part of the LISA framework, bridges gaps identified in current screening practices, offering a systematic and user-friendly tool that would allow for a seamless transition from universal identification of students' needs to the implementation of learning strategies. The current government support and recognition, as evidenced by the substantial grant for a widespread integration plan, position LISA as a transformative initiative within the French education system.

Overall, the work described in the thesis contributes to improvement of identification and management of mental health and learning disorder in children by, on one hand, providing researchers with a data-driven framework that simplifies the creation of new assessments, and on the other hand, providing schools with tools to systematically screen for, and address early signs of mental and learning disorders.

These two approaches can complement each other effectively. The framework described in Chapter 2 can be used to create disorder-specific screeners, which then will be systematically administered as a part of the LISA framework to children with concerns identified by FACETS that are consistent with specific disorders. The framework can also be used to create diagnostic assessments used in the clinic, where children will be referred to through the systematic referral process currently being implemented within the LISA platform.

4.2. Perspectives

Besides the future directions outlined in the respective discussion sections of each chapter, further application of the results obtained in this thesis can be envisaged.

An integration of multi-label classification with item response theory methods into the item-recommender package, as described in Gibbons et al. (2016), can be used to create a computerized adaptive diagnostic test. Probability calibration techniques can be used to estimate the probability of each disorder, instead of a binary diagnosis. A self-report and parent-report version of the test can be accompanied by recommendations of evidence-based resources and self- and caregiver-administered interventions relevant to the diagnosis, similar to the approach of the LISA framework and the CMI Symptom Checker (*Symptom Checker*, n.d.).

A web interface can be developed that would allow researchers to consult items that have been shown to be predictive of a particular disorder, for building item banks for new assessments. This can be supplemented by semantic similarity analysis, where the text of the items is matched with a description of a disorder or an evaluated domain (e.g. Attention) by their semantic content, independent of responses to the items. This approach is useful in cases where not enough response labeled response data is available. Initial semantic similarity analysis of HBN items was conducted by Kai McClennen as a part of his internship this summer at the Senseable Intelligence Group,

where efficiency of different semantic representation models was examined. In addition to the HBN items, the items can be extracted from the linked open database of resources related to mental health (*Mhdb*, 2020/2021) that has been built by the MATTER Lab, under the direction of Arno Klein. The database contains thousands of items from established mental health questionnaires, including items used in the HBN study.

The semantic similarity approach can be further used for matching all HBN items to a set of dimensions (e.g. continuous FACETS items) and defining a mapping function between item response options and location on the dimensions to create dimensional profiles of HBN participants. Cluster analysis can be applied to the dimensional profiles to explore trans-diagnostic groups among participants. These groups could be compared to existing diagnostic categories, or be used to explore potential correlates with genetic and/or neuroimaging data.

Large scale studies incorporating FACETS responses and effectiveness of classroom or individual-level interventions will provide a dataset that can be used for building a machine-learning based recommendation system of interventions that have been shown to be effective for students with similar profiles in the past.

Additionally, clinicians at the Robert-Debré hospital expressed interest in FACETS as a potential initial screening tool to be administered to every child entering the child psychiatry department. Further evaluation of validity of FACETS and investigation of its potential use as a screening tool in clinical settings is required to confirm its utility for this purpose. If the studies that are currently underway show good psychometric properties of FACETS, it can be potentially integrated into the protocol of larger studies, such as HBN or other similar initiatives, to obtain further psychometric support.

In the process of development of LISA-DB, the database of guides and strategies for teachers and parents, the Mosaic project emerged, seeking to address the fragmented landscape of youth mental health resources. During the development of the database, we encountered a body of disconnected articles and guides with inconsistent quality and relevance, often not updated regularly. The Mosaic project aims to centralize these resources efficiently, creating a unified, multilingual, curated, high-quality database of mental health resources. The project aims to develop a rich schema accommodating a spectrum of mental health resource repositories, including LISA-DB, CMI Resource Center (*Family Resource Center*, n.d.), CAMHI guides (*Clinical Short Guides CAMHI*, n.d.), and CléPsy guides (*CléPsy*, n.d.).

Mosaic aims to serve as a comprehensive resource hub, offering a single source of truth for different clients/products within the youth mental health domain. The database, its website, and APIs will facilitate queries targeting various end-users and contexts, such as clinicians, researchers, educators, and parents/caregivers. Clinicians can benefit from the database by being able to access assessments, interventions, training guides, and foundational elements such as signs and symptoms of DSM disorders. Researchers will be able to use Mosaic in their research by being able to query the relationships encoded in the database, such as links between objectively observable/measurable phenomena and mental constructs. Educators will have access to behavioral assessments and behavior management guides, while parents/caregivers can use the database for informational resources and guides containing advice relevant to their support roles.

In the last years, other initiatives emerged, focusing on promotion of mental health among school-age children.

Schools4Health (*Schools4Health - Schools for Health*, n.d.), spanning from 2023 to 2025, aims to establish and fortify a comprehensive, participatory approach to health and well-being within schools in Europe. Focused on nutrition, physical activity, and mental health, the initiative intends to showcase the effectiveness of Health Promoting Schools (HPS) and other whole-school strategies. Its objectives encompass promoting and implementing HPS approaches, influencing policymakers, identifying and implementing best practices for healthy lifestyles, and emphasizing the broader contributions of HPS to equity and the environment. The LISA framework aligns with the goals of Schools4Health, offering a whole-school approach to identify and implement evidence based assessment and intervention. LISA's capacity to streamline assessments, offer personalized educational plans, and provide ongoing support to teachers and parents aligns with the participatory nature of Schools4Health, contributing to creating healthier school environments.

The Child Mind Institute, in partnership with the Stavros Niarchos Foundation (SNF), is launching the Global Center for Child and Adolescent Mental Health (*The Stavros Niarchos Foundation Global Center for Child and Adolescent Mental Health*, n.d.). This initiative is committed to advancing global collaboration in the often under-researched domains of children's mental health, with a vision of providing easily accessible, high-quality information, resources, and care for children and families worldwide. The primary focus revolves around breaking down barriers to mental health care by expanding service access, disseminating accessible information, and combating mental health stigma. Integrated with the LISA framework, this initiative could leverage LISA's capabilities to

enhance collaboration between mental health experts to build evidence-based guides, streamline information dissemination, and offer targeted training to teachers on evidence-based interventions.

UPRIGHT (“Upright,” n.d.) and BOOST (*Boost Project | Promoting Mental Health Resilience*, n.d.) and two research initiatives funded by the European Union's HORIZON 2020 program who aim to cultivate a mental well-being culture in schools across five European regions. BOOST focuses primarily on social and emotional learning in primary school, while UPRIGHT focuses on secondary school students and takes a broader approach with a focus on coping, efficacy, social and emotional learning, and mindfulness. The projects aim to empower students to apply resilience skills in their daily lives, thereby enhancing their well-being and that of their families. UPRIGHT and BOOST training programs are in line with the goals of LISA to improve students’ wellbeing on the school level. Such training programs will become a part of the LISA training programs for teachers, thus being integrating into the educational process in France through the systematic adoption of LISA within the French public education system.

Bibliography

Abidin, R. R. (2012). Parenting stress index–fourth edition (PSI-4). *Lutz, FL: Psychological Assessment Resources*.

Aboraya, A. (2007). The Reliability of Psychiatric Diagnoses: Point—Our psychiatric Diagnoses are Still Unreliable. *Psychiatry (Edgmont)*, *4*(1), 22.

Achenbach, T. M. (1991). *Integrative Guide for the 1991 CBCL/4-18, Ysr, and Trf Profiles* (1st US-1st Printing edition). Univ Vermont/Dept Psychiatry.

Achenie, L. E. K., Scarpa, A., Factor, R. S., Wang, T., Robins, D. L., & McCrickard, D. S. (2019). A Machine Learning Strategy for Autism Screening in Toddlers. *Journal of Developmental and Behavioral Pediatrics : JDBP*, *40*(5), 369–376. https://doi.org/10.1097/DBP.0000000000000668

Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., Litke, S., O'Hagan, B., Andersen, J., Bronstein, B., Bui, A., Bushey, M., Butler, H., Castagna, V., Camacho, N., … Milham, M. P. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, *4*(1), Article 1. https://doi.org/10.1038/sdata.2017.181

Alexander, L., Salum, G., Milham, M., & Swanson, J. (n.d.). *Extended Strengths and Weaknesses Assessment of Normal Behavior (E-SWAN)*. Retrieved November 16, 2023, from http://www.eswan.org/

Algamal, Z. Y., & Lee, M. H. (2015). Applying Penalized Binary Logistic Regression with Correlation Based Elastic Net for Variables Selection. *Journal of Modern Applied Statistical Methods*, *14*(1), 168–179. https://doi.org/10.22237/jmasm/1430453640

Allgaier, A.-K., Pietsch, K., Frühe, B., Sigl-Glöckner, J., & Schulte-Körne, G. (2012). Screening for depression in adolescents: Validity of the patient health questionnaire in pediatric care. *Depression and Anxiety*, *29*(10), 906–913. https://doi.org/10.1002/da.21971

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Anderson, J. S., Shade, J., DiBlasi, E., Shabalin, A. A., & Docherty, A. R. (2019). Polygenic risk scoring and prediction of mental health outcomes. *Current Opinion in Psychology*, *27*, 77–81. https://doi.org/10.1016/j.copsyc.2018.09.002

Anderson, M., Werner-Seidler, A., King, C., Gayed, A., Harvey, S. B., & O'Dea, B. (2019). Mental Health Training Programs for Secondary School Teachers: A Systematic Review. *School Mental Health*, *11*(3), 489–508. https://doi.org/10.1007/s12310-018-9291-2

Angold, A., Costello, E., Messer, S., Pickles, A., Winder, F., & Silver, D. (1995). The development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *Int J Methods Psychiatr Res*, *5*, 1–12.

Angold, A., Erkanli, A., Farmer, E. M. Z., Fairbank, J. A., Burns, B. J., Keeler, G., & Costello, E. J. (2002). Psychiatric Disorder, Impairment, and Service Use in Rural African American and White Youth. *Archives of General Psychiatry*, *59*(10), 893. https://doi.org/10.1001/archpsyc.59.10.893

Angst, J. (2007). The bipolar spectrum. *British Journal of Psychiatry*, *190*(3), 189–191. https://doi.org/10.1192/bjp.bp.106.030957

Armitage, J. M., Tseliou, F., Riglin, L., Dennison, C., Eyre, O., Jones, R. B., Rice, F., Thapar, A. K., Thapar, A., & Collishaw, S. (2023). Validation of the Strengths and Difficulties Questionnaire (SDQ) emotional subscale in assessing depression and anxiety across development. *PLOS ONE*, *18*(7), e0288882. https://doi.org/10.1371/journal.pone.0288882

Ayers, T., Sandler, I., Bernzweig, J., Harrison, R., Wampler, T., & Lustig, J. (1989). Handbook for the content analysis of children's coping responses. *Tempe (AZ): Program for Prevention Research, Arizona State University*.

Baak, M., Miller, E., Ziersch, A., Due, C., Masocha, S., & Ziaian, T. (2020). The Role of Schools in Identifying and Referring Refugee Background Young People Who Are Experiencing Mental Health Issues. *Journal of School Health*, *90*(3), 172–181. https://doi.org/10.1111/josh.12862

Balogh, E. P., Miller, B. T., Ball, J. R., Care, C. on D. E. in H., Services, B. on H. C., Medicine, I. of, & The National Academies of Sciences, E. (2015). The Diagnostic Process. In *Improving Diagnosis in Health Care*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK338593/

Bard, D. E., Wolraich, M. L., Neas, B., Doffing, M., & Beck, L. (2013). The Psychometric Properties of the Vanderbilt Attention-Deficit Hyperactivity Disorder Diagnostic Parent Rating Scale in a Community Population. *Journal of Developmental & Behavioral Pediatrics*, *34*(2), 72. https://doi.org/10.1097/DBP.0b013e31827a3a22

Barratt, W. (2012, June 14). The Barratt Simplified Measure of Social Status (BSMSS). *Social Class on Campus*. http://socialclassoncampus.blogspot.com/2012/06/barratt-simplified-measure-of-social.html

Bartosik, A., & Whittingham, H. (2021). Chapter 7—Evaluating safety and toxicity. In S. K. Ashenden (Ed.), *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry* (pp. 119–137). Academic Press. https://doi.org/10.1016/B978-0-12-820045-2.00008-8

Becker-Haimes, E. M., Tabachnick, A. R., Last, B. S., Stewart, R. E., Hasan-Granier, A., & Beidas, R. S. (2020). Evidence Base Update for Brief, Free, and Accessible Youth Mental Health Measures. *Journal of Clinical Child and Adolescent Psychology : The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *49*(1), 1–17. https://doi.org/10.1080/15374416.2019.1689824

Bird, H. R., Shaffer, D., Fisher, P., Gould, M. S., & et al. (1993). The Columbia Impairment Scale (CIS): Pilot findings on a measure of global impairment for children and adolescents. *International Journal of Methods in Psychiatric Research*, *3*(3), 167–176.

Birmaher, B., Khetarpal, S., Brent, D., Cully, M., Balach, L., Kaufman, J., & Neer, S. M. (1997). The Screen for Child Anxiety Related Emotional Disorders (SCARED): Scale construction and psychometric characteristics. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*(4), 545–553. https://doi.org/10.1097/00004583-199704000-00018

Bjork, J. M., Straub, L. K., Provost, R. G., & Neale, M. C. (2017). The ABCD study of neurodevelopment: Identifying neurocircuit targets for prevention and treatment of

adolescent substance abuse. *Current Treatment Options in Psychiatry*, *4*(2), 196–209. https://doi.org/10.1007/s40501-017-0108-y

Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., Lord, C., & Narayanan, S. S. (2016). Use of machine learning to improve autism screening and diagnostic instruments: Effectiveness, efficiency, and multi-instrument fusion. *Journal of Child Psychology and Psychiatry*, *57*(8), 927–937. https://doi.org/10.1111/jcpp.12559

Bone, D., Goodwin, M. S., Black, M. P., Lee, C.-C., Audhkhasi, K., & Narayanan, S. (2015). Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders*, *45*(5), 1121–1136. https://doi.org/10.1007/s10803-014-2268-6

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika*, *65*(1), 23–28. https://doi.org/10.1007/BF02294183

Boost Project | promoting mental health resilience. (n.d.). Retrieved November 15, 2023, from https://www.boostproject.eu/

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Botchkarev, A. (2019). Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. *Interdisciplinary Journal of Information, Knowledge, and Management*, *14*, 045–076. https://doi.org/10.28945/4184

Braaten, E. B., Ward, A. K., Forchelli, G., Vuijk, P. J., Cook, N. E., McGuinness, P., Lee, B. A., Samkavitz, A., Lind, H., O'Keefe, S. M., & Doyle, A. E. (2020). Characteristics of child psychiatric outpatients with slow processing speed and potential mechanisms of academic impact. *European Child & Adolescent Psychiatry*, *29*(10), 1453–1464. https://doi.org/10.1007/s00787-019-01455-w

Brauner, C. B., & Stephens, C. B. (2006). Estimating the Prevalence of Early Childhood Serious Emotional/Behavioral Disorders: Challenges and Recommendations. *Public Health Reports*, *121*(3), 303–310. https://doi.org/10.1177/003335490612100314

Brodey, B. B., Girgis, R. R., Favorov, O. V., Addington, J., Perkins, D. O., Bearden, C. E., Woods, S. W., Walker, E. F., Cornblatt, B. A., Brucato, G., Walsh, B., Elkin, K. A., & Brodey, I. S. (2018). The Early Psychosis Screener (EPS): Quantitative validation against the SIPS using machine learning. *Schizophrenia Research*, *197*, 516–521. https://doi.org/10.1016/j.schres.2017.11.030

Brückl, M. (2018). *irrNA: Coefficients of Interrater Reliability – Generalized for Randomly Incomplete Datasets*.

Bruhn, A. L., Woods-Groves, S., & Huddle, S. (2014). A preliminary investigation of emotional and behavioral screening practices in K–12 schools. *Education & Treatment of Children*, *37*(4), 611–634. https://doi.org/10.1353/etc.2014.0039

Bruno, P. (2011). The importance of diagnostic test parameters in the interpretation of clinical test findings: The Prone Hip Extension Test as an example. *The Journal of the Canadian Chiropractic Association*, *55*(2), 69–75.

Bruns, E. J., Duong, M. T., Lyon, A. R., Pullmann, M. D., Cook, C. R., Cheney, D., & McCauley, E. (2016). Fostering SMART Partnerships to Develop an Effective Continuum of Behavioral Health Services and Supports in Schools. *The American Journal of Orthopsychiatry*, *86*(2), 156–170. https://doi.org/10.1037/ort0000083

Bujang, M. A. (2017). A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. *Archives of Orofacial Sciences*, *12*, 1–11.

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of Classical Test Theory and Item Response Theory for Quantitative Assessment of Items in Developing Patient-Reported Outcome Measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Carpenter, K. L. H., Sprechmann, P., Calderbank, R., Sapiro, G., & Egger, H. L. (2016). Quantifying Risk for Anxiety Disorders in Preschool Children: A Machine Learning

Approach. *PloS One*, *11*(11), e0165524. https://doi.org/10.1371/journal.pone.0165524

Casat, C. D., Norton, H. J., & Boyle-Whitesel, M. (1999). Identification of Elementary School Children at Risk for Disruptive Behavioral Disturbance: Validation of a Combined Screening Method. *Journal of the American Academy of Child & Adolescent Psychiatry*, *38*(10), 1246–1253. https://doi.org/10.1097/00004583-199910000-00013

Charman, T., & Gotham, K. (2013). Measurement Issues: Screening and diagnostic instruments for autism spectrum disorders - lessons from research and practise. *Child and Adolescent Mental Health*, *18*(1), 52–63. https://doi.org/10.1111/j.1475-3588.2012.00664.x

Chawarska, K., Klin, A., & Volkmar, F. R. (2008). *Autism Spectrum Disorders in Infants and Toddlers: Diagnosis, Assessment, and Treatment*. Guilford Press.

Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2020). *A Survey of Model Compression and Acceleration for Deep Neural Networks* (arXiv:1710.09282). arXiv. http://arxiv.org/abs/1710.09282

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ. Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Clark, L. A., Cuthbert, B., Lewis-Fernández, R., Narrow, W. E., & Reed, G. M. (2017). Three Approaches to Understanding and Classifying Mental Disorder: ICD-11, DSM-5, and the National Institute of Mental Health's Research Domain Criteria (RDoC). *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *18*(2), 72–145. https://doi.org/10.1177/1529100617727266

CléPsy. (n.d.). CléPsy. Retrieved November 14, 2023, from https://www.clepsy.fr/

Clinical short guides CAMHI. (n.d.). CAMHI. Retrieved November 14, 2023, from https://camhi.gr/en/health-professionals/clinical-short-guides/

Colizzi, M., Lasalvia, A., & Ruggeri, M. (2020). Prevention and early intervention in youth mental health: Is it time for a multidisciplinary and trans-diagnostic model for care? *International Journal of Mental Health Systems*, *14*(1), 23. https://doi.org/10.1186/s13033-020-00356-9

Conners, C. K. (2008). *Conners 3rd Edition*. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Behavior/Comprehensive/Conners-3rd-Edition/p/100000523.html

Connors, E. H., Moffa, K., Carter, T., Crocker, J., Bohnenkamp, J. H., Lever, N. A., & Hoover, S. A. (2022). Advancing Mental Health Screening in Schools: Innovative, Field-Tested Practices and Observed Trends During a 15-Month Learning Collaborative. *Psychology in the Schools*, *59*(6), 1135–1157. https://doi.org/10.1002/pits.22670

Constantino, J. N., Davis, S. A., Todd, R. D., Schindler, M. K., Gross, M. M., Brophy, S. L., Metzger, L. M., Shoushtari, C. S., Splinter, R., & Reich, W. (2003). Validation of a brief quantitative measure of autistic traits: Comparison of the social responsiveness scale with the autism diagnostic interview-revised. *Journal of Autism and Developmental Disorders*, *33*(4), 427–433. https://doi.org/10.1023/a:1025014929212

Costa, D. S., Paula, J. J. de, Malloy-Diniz, L. F., Romano-Silva, M. A., & Miranda, D. M. (2019). Parent SNAP-IV rating of attention-deficit/hyperactivity disorder: Accuracy in a clinical sample of ADHD, validity, and reliability in a Brazilian sample. *Jornal de Pediatria*, *95*, 736–743. https://doi.org/10.1016/j.jped.2018.06.014

Cullinan, V., Veale, A., & Vitale, A. (2016). Irish General Practitioner referrals to psychological therapies. *Irish Journal of Psychological Medicine*, *33*(2), 73–80. https://doi.org/10.1017/ipm.2015.17

Curtin, J. J. (2020). *Introduction to Applied Machine Learning*. https://dionysus.psych.wisc.edu/iaml_2020/

Dagani, J., Signorini, G., Nielssen, O., Bani, M., Pastore, A., de Girolamo, G., & Large, M. (2017). Meta-analysis of the Interval between the Onset and Management of Bipolar Disorder. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *62*(4), 247–258. https://doi.org/10.1177/0706743716656607

Dalsgaard, S., McGrath, J., Østergaard, S. D., Wray, N. R., Pedersen, C. B., Mortensen, P. B., & Petersen, L. (2020). Association of Mental Disorder in Childhood and Adolescence With Subsequent Educational Achievement. *JAMA Psychiatry*, *77*(8), 797–805. https://doi.org/10.1001/jamapsychiatry.2020.0217

Dattani, S. (2023, May 26). *How do researchers study the prevalence of mental illnesses?* Our World in Data. https://ourworldindata.org/how-do-researchers-study-the-prevalence-of-mental-illnesses

Daveney, J., Panagioti, M., Waheed, W., & Esmail, A. (2019). Unrecognized bipolar disorder in patients with depression managed in primary care: A systematic review and meta-analysis. *General Hospital Psychiatry*, *58*, 71–76. https://doi.org/10.1016/j.genhosppsych.2019.03.006

De Girolamo, G., Dagani, J., Purcell, R., Cocchi, A., & McGorry, P. D. (2012). Age of onset of mental disorders and use of mental health services: Needs, opportunities and obstacles. *Epidemiology and Psychiatric Sciences*, *21*(1), 47–57. https://doi.org/10.1017/S2045796011000746

de Graaf, R., Kessler, R. C., Fayyad, J., ten Have, M., Alonso, J., Angermeyer, M., Borges, G., Demyttenaere, K., Gasquet, I., de Girolamo, G., Haro, J. M., Jin, R., Karam, E. G., Ormel, J., & Posada-Villa, J. (2008). The prevalence and effects of adult attention-deficit/hyperactivity disorder (ADHD) on the performance of workers: Results from the WHO World Mental Health Survey Initiative. *Occupational and Environmental Medicine*, *65*(12), 835–842. https://doi.org/10.1136/oem.2007.038448

de Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory Factor Analysis With Small Sample Sizes. *Multivariate Behavioral Research*, *44*(2), 147–181. https://doi.org/10.1080/00273170902794206

Dekel, O., & Shamir, O. (2010). Multiclass-Multilabel Classification with More Classes than Examples. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 137–144. https://proceedings.mlr.press/v9/dekel10a.html

Dharma, C., Grace, D., Logie, C., Abramovich, A., Mitsitakis, N., Baskerville, N., & Chaiton, M. (2022). *Identifying Factors Affecting Depressive Symptoms and Incidence of Mental Health Diagnosis within 1 Year among 2SLGBTQ+ Youth*

During COVID-19 Using Machine Learning Methods. https://doi.org/10.21203/rs.3.rs-2199889/v1

Dombrowski, S. C. (Ed.). (2020). *Psychoeducational Assessment and Report Writing*. Springer International Publishing. https://doi.org/10.1007/978-3-030-44641-3

drivendata/cookiecutter-data-science: A logical, reasonably standardized, but flexible project structure for doing and sharing data science work. (n.d.). Retrieved November 16, 2023, from https://github.com/drivendata/cookiecutter-data-science

Drobatz, K. J. (2009). Measures of accuracy and performance of diagnostic tests. *Journal of Veterinary Cardiology*, *11*, S33–S40. https://doi.org/10.1016/j.jvc.2009.03.004

Duncan, L., Georgiades, K., Wang, L., Edwards, J., & Comeau, J. (2022). Estimating prevalence of child and youth mental disorder and mental health-related service contacts: A comparison of survey data and linked administrative health data. *Epidemiology and Psychiatric Sciences*, *31*, e35. https://doi.org/10.1017/S204579602200018X

Dwyer, S. B., Nicholson, J. M., & Battistutta, D. (2006). Parent and Teacher Identification of Children at Risk of Developing Internalizing or Externalizing Mental Health Problems: A Comparison of Screening Methods. *Prevention Science*, *7*(4), 343–357. https://doi.org/10.1007/s11121-006-0026-5

École et handicap—PPS, PAI, PAP, PPRE. (2021, September 29). Mon Parcours Handicap. https://www.monparcourshandicap.gouv.fr/scolarite/ppre-pai-pap-pps-en-quoi-consistent-les-differentes-possibilites-dappui-la-scolarisation

École et handicap—Qu'est-ce que le GEVA-sco ? (2021, September 29). Mon Parcours Handicap. https://www.monparcourshandicap.gouv.fr/scolarite/quest-ce-que-le-geva-sco

Ehlers, S., & Gillberg, C. (1993). The epidemiology of Asperger syndrome. A total population study. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *34*(8), 1327–1350. https://doi.org/10.1111/j.1469-7610.1993.tb02094.x

Eleje, L. I., Onah, F. E., & Abanobi, C. C. (n.d.). *Comparative Study of Classical Test Theory and Item Response Theory Using Diagnostic Quantitative Economics Skill Test Item Analysis Results*.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), 140. https://doi.org/10.1186/s40537-021-00516-9

Enabee – étude nationale sur le bien être des enfants. (n.d.). Retrieved November 14, 2023, from https://enabee.fr/

Enøe, C., Georgiadis, M. P., & Johnson, W. O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine*, *45*(1), 61–81. https://doi.org/10.1016/S0167-5877(00)00117-3

Essau, C. A., Sasagawa, S., & Frick, P. J. (2006a). Callous-unemotional traits in a community sample of adolescents. *Assessment*, *13*(4), 454–469. https://doi.org/10.1177/1073191106287354

Essau, C. A., Sasagawa, S., & Frick, P. J. (2006b). Psychometric Properties of the Alabama Parenting Questionnaire. *Journal of Child and Family Studies*, *15*(5), 597–616. https://doi.org/10.1007/s10826-006-9036-y

Étape 1: Je compose ma grille d'observation. (n.d.). Retrieved November 15, 2023, from https://www.reseau-canope.fr/cap-ecole-inclusive/observer.html?tx_cndphandicap_cndphandicap%5Bniveau%5D=1&tx_cndphandicap_cndphandicap%5Baction%5D=createGrille&tx_cndphandicap_cndphandicap%5Bcontroller%5D=Grille&cHash=1d3f2e608425f374823c3d9ffd78116e

Falkmer, T., Anderson, K., Falkmer, M., & Horlin, C. (2013). Diagnostic procedures in autism spectrum disorders: A systematic literature review. *European Child & Adolescent Psychiatry*, *22*(6), 329–340. https://doi.org/10.1007/s00787-013-0375-0

Family Resource Center. (n.d.). Child Mind Institute. Retrieved November 14, 2023, from https://childmind.org/resources/

Farrell, L. J., & Barrett, P. M. (2007). Prevention of Childhood Emotional Disorders: Reducing the Burden of Suffering Associated with Anxiety and Depression. *Child and Adolescent Mental Health*, *12*(2), 58–65. https://doi.org/10.1111/j.1475-3588.2006.00430.x

Fazel, M., Hoagwood, K., Stephan, S., & Ford, T. (2014). Mental health interventions in schools in high-income countries. *The Lancet Psychiatry*, *1*(5), 377–387. https://doi.org/10.1016/S2215-0366(14)70312-8

Feng, J., Wang, Y., Peng, J., Sun, M., Zeng, J., & Jiang, H. (2019). Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. *Journal of Critical Care*, *54*, 110–116. https://doi.org/10.1016/j.jcrc.2019.08.010

Ferenchick, E. K., Ramanuj, P., & Pincus, H. A. (2019). Depression in primary care: Part 1—screening and diagnosis. *BMJ*, l794. https://doi.org/10.1136/bmj.l794

Fernández, A., Rubio-Valera, M., Bellón, J. A., Pinto-Meza, A., Luciano, J. V., Mendive, J. M., Haro, J. M., Palao, D. J., & Serrano-Blanco, A. (2012). Recognition of anxiety disorders by the general practitioner: Results from the DASMAP Study. *General Hospital Psychiatry*, *34*(3), 227–233. https://doi.org/10.1016/j.genhosppsych.2012.01.012

Ferrari, A. J., Somerville, A. J., Baxter, A. J., Norman, R., Patten, S. B., Vos, T., & Whiteford, H. A. (2013). Global variation in the prevalence and incidence of major depressive disorder: A systematic review of the epidemiological literature. *Psychological Medicine*, *43*(3), 471–481. https://doi.org/10.1017/S0033291712001511

Fischer, J. E., Bachmann, L. M., & Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: Clinical example of sepsis. *Intensive Care Medicine*, *29*(7), 1043–1051. https://doi.org/10.1007/s00134-003-1761-8

Fisher, R. A. (1992). Statistical Methods for Research Workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Methodology and Distribution* (pp. 66–70). Springer. https://doi.org/10.1007/978-1-4612-4380-9_6

Flake, J. K., & Fried, E. I. (2019). *Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them*. PsyArXiv. https://doi.org/10.31234/osf.io/hs7wm

Forlin, C., & Chambers, D. (2011). Teacher preparation for inclusive education: Increasing knowledge but raising concerns. *Asia-Pacific Journal of Teacher Education*, *39*, 17–32. https://doi.org/10.1080/1359866X.2010.540850

Fried, E. I., & Flake, J. K. (2018). Measurement Matters. *APS Observer*, *31*. https://www.psychologicalscience.org/observer/measurement-matters

Furukawa, T. A., Strauss, S. E., Bucher, H. C., Thomas, A., & Guyatt, G. (2015). Diagnostic Tests. In G. Guyatt, D. Rennie, M. O. Meade, & D. J. Cook (Eds.), *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 3rd ed* (1–Book, Section). McGraw-Hill Education. jamaevidence.mhmedical.com/content.aspx?aid=1183877151

Fusar-Poli, P., Correll, C. U., Arango, C., Berk, M., Patel, V., & Ioannidis, J. P. A. (2021). Preventive psychiatry: A blueprint for improving the mental health of young people. *World Psychiatry*, *20*(2), 200–221. https://doi.org/10.1002/wps.20869

Gaab, N., & Petscher, Y. (2022). Screening for Early Literacy Milestones and Reading Disabilities. *EarlyBird Education*. https://earlybirdeducation.com/educators-posts/screening-for-early-literacy-milestones-and-reading-disabilities/

Gaab, N., Turesky, T. K., & Sanfilippo, J. (n.d.). *Early Identification of Children at Risk for Reading Difficulty*.

Gaebel, W., Stricker, J., & Kerst, A. (2020). Changes from ICD-10 to ICD-11 and future directions in psychiatric classification. *Dialogues in Clinical Neuroscience*, *22*(1), 7–15. https://doi.org/10.31887/DCNS.2020.22.1/wgaebel

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* (0.84.1) [Computer software]. https://cran.r-project.org/web/packages/irr/index.html

Ganguly, S., Samanta, M., Roy, P., Chatterjee, S., Kaplan, D. W., & Basu, B. (2013). Patient Health Questionnaire-9 as an Effective Tool for Screening of Depression Among Indian Adolescents. *Journal of Adolescent Health*, *52*(5), 546–551. https://doi.org/10.1016/j.jadohealth.2012.09.012

Gater, R., Almeida E. Sousa, D. B., Barrientos, G., Caraveo, J., Chandrashekar, C. R., Dhadphale, M., Goldberg, D., Al Kathiri, A. H., Mubbashar, M., Silhan, K., Thong, D., Torres-Gonzales, F., & Sartorius, N. (1991). The pathways to psychiatric care: A cross-cultural study. *Psychological Medicine*, *21*(3), 761–774. https://doi.org/10.1017/S003329170002239X

Geisinger, K. F., Bracken, B. A., Carlson, J. F., Hansen, J.-I. C., Kuncel, N. R., Reise, S. P., & Rodriguez, M. C. (Eds.). (2013). *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology.* American Psychological Association. https://doi.org/10.1037/14047-000

Gibbons, R. D., Chattopadhyay, I., Meltzer, H., Kane, J. M., & Guinart, D. (2022). Development of a Computerized Adaptive Diagnostic Screening Tool for Psychosis. *Schizophrenia Research*, *245*, 116–121. https://doi.org/10.1016/j.schres.2021.03.020

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized Adaptive Diagnosis and Testing of Mental Health Disorders. *Annual Review of Clinical Psychology*, *12*, 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634

Gibbs, S., & Miller, A. (2014). Teachers' resilience and well-being: A role for educational psychology. *Teachers and Teaching*, *20*(5), 609–621. https://doi.org/10.1080/13540602.2013.844408

Ginsberg, Y., Quintero, J., Anand, E., Casillas, M., & Upadhyaya, H. P. (2014). Underdiagnosis of Attention-Deficit/Hyperactivity Disorder in Adult Patients: A Review of the Literature. *The Primary Care Companion for CNS Disorders*, *16*(3), 23591. https://doi.org/10.4088/PCC.13r01600

Glavin, D., Grua, E. M., Nakamura, C. A., Scazufca, M., Santos, E. R. dos, Wong, G. H. Y., Hollingworth, W., Peters, T. J., Araya, R., & Ven, P. V. de. (2023). Patient Health Questionnaire-9 Item Pairing Predictiveness for Prescreening Depressive Symptomatology: Machine Learning Analysis. *JMIR Mental Health*, *10*(1), e48444. https://doi.org/10.2196/48444

Gleason, P. M., Harris, J., Sheean, P. M., Boushey, C. J., & Bruemmer, B. (2010). Publishing Nutrition Research: Validity, Reliability, and Diagnostic Test Assessment in Nutrition-Related Research. *Journal of the American Dietetic Association*, *110*(3), 409–419. https://doi.org/10.1016/j.jada.2009.11.022

Gold, L. H. (2014). DSM-5 and the assessment of functioning: The World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0). *The Journal of the American Academy of Psychiatry and the Law*, *42*(2), 173–181.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *38*(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2003). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *International Review of Psychiatry*, *15*(1–2), 166–172. https://doi.org/10.1080/0954026021000046128

Graham, A., Phelps, R., Maddison, C., & Fitzgerald, R. (2011). Supporting children's mental health in schools: Teacher views. *Teachers and Teaching*, *17*(4), 479–496. https://doi.org/10.1080/13540602.2011.580525

Gray, C., Wilcox, G., & Nordstokke, D. (2017). Teacher Mental Health, School Climate, Inclusive Education and Student Learning: A Review. *Canadian Psychology/Psychologie Canadienne*, *58*, 203–210. https://doi.org/10.1037/cap0000117

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Gudivada, V. N., Irfan, M. T., Fathi, E., & Rao, D. L. (2016). Chapter 5 - Cognitive Analytics: Going Beyond Big Data Analytics and Machine Learning. In V. N. Gudivada, V. V. Raghavan, V. Govindaraju, & C. R. Rao (Eds.), *Handbook of*

Statistics (Vol. 35, pp. 169–205). Elsevier. https://doi.org/10.1016/bs.host.2016.07.010

Guo, S., & Jhe, G. B. (2021). Universal Depression Screening in Schools—Promises and Challenges in Addressing Adolescent Mental Health Need. *JAMA Network Open*, *4*(11), e2132858. https://doi.org/10.1001/jamanetworkopen.2021.32858

Hägele, C., Schlagenhauf, F., Rapp, M., Sterzer, P., Beck, A., Bermpohl, F., Stoy, M., Ströhle, A., Wittchen, H.-U., Dolan, R. J., & Heinz, A. (2015). Dimensional psychiatry: Reward dysfunction and depressive mood across psychiatric disorders. *Psychopharmacology*, *232*(2), 331–341. https://doi.org/10.1007/s00213-014-3662-7

Hagell, A., Wortley, L., Ross, L., & Whitaker, E. (2021). *Key Data on Young People 2021 Overview and Policy Implications*. https://www.ayph-youthhealthdata.org.uk/

Hagen, T. O., Kristine Amlund. (2018). *Adolescent Mental Health: Prevention and Intervention* (2nd ed.). Routledge. https://doi.org/10.4324/9781315295374

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3e édition). Morgan Kaufmann Publishers In.

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*(1), 167–178. https://doi.org/10.1037/0033-2909.117.1.167

Hayes, A. M., Dombrowski, E., Shefcyk, A., & Bulat, J. (2018). *Learning Disabilities Screening and Evaluation Guide for Low- and Middle-Income Countries*. RTI Press. http://www.ncbi.nlm.nih.gov/books/NBK545498/

He, J.-P., Burstein, M., Schmitz, A., & Merikangas, K. R. (2013). The Strengths and Difficulties Questionnaire (SDQ): The Factor Structure and Scale Validation in U.S. Adolescents. *Journal of Abnormal Child Psychology*, *41*(4), 583–595. https://doi.org/10.1007/s10802-012-9696-6

Hetland, J., Braatveit, K. J., Hagen, E., Lundervold, A. J., & Erga, A. H. (2021). Prevalence and Characteristics of Borderline Intellectual Functioning in a Cohort of Patients With Polysubstance Use Disorder. *Frontiers in Psychiatry*, *12*, 651028. https://doi.org/10.3389/fpsyt.2021.651028

Hirsch, J. A., Nicola, G., McGinty, G., Liu, R. W., Barr, R. M., Chittle, M. D., & Manchikanti, L. (2016). ICD-10: History and Context. *AJNR: American Journal of Neuroradiology*, *37*(4), 596–599. https://doi.org/10.3174/ajnr.A4696

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.2307/1267351

Institut national de la santé et de la recherche Institut national de la santé et de la recherche. (2002). Troubles mentaux: Dépistage et prévention chez l'enfant et

l'adolescent. *Collection Expertise collective Inserm*. https://www.ipubli.inserm.fr/handle/10608/165

Je souhaite m'engager dans la démarche École promotrice de santé. (2023, juillet). éduscol | Ministère de l'Éducation nationale et de la Jeunesse - Direction générale de l'enseignement scolaire. https://eduscol.education.fr/2063/je-souhaite-m-engager-dans-la-demarche-ecole-promotrice-de-sante

Jellinek, M. S., Murphy, J. M., Robinson, J., Feins, A., Lamb, S., & Fenton, T. (1988). Pediatric Symptom Checklist: Screening school-age children for psychosocial dysfunction. *The Journal of Pediatrics*, *112*(2), 201–209. https://doi.org/10.1016/S0022-3476(88)80056-8

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, *51*(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Johnson, C., Eva, A. L., Johnson, L., & Walker, B. (2011). Don't Turn Away: Empowering Teachers to Support Students' Mental Health. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *84*(1), 9–14. https://doi.org/10.1080/00098655.2010.484441

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How Can We Improve the Accuracy of Screening Instruments? *Learning Disabilities Research & Practice*, *24*(4), 174–185. https://doi.org/10.1111/j.1540-5826.2009.00291.x

Johnston, O. G., & Burke, J. D. (2020). Parental Problem Recognition and Help-Seeking for Disruptive Behavior Disorders. *The Journal of Behavioral Health Services & Research*, *47*(1), 146–163. https://doi.org/10.1007/s11414-018-09648-y

Jones, P. B. (2013). Adult mental health disorders and their age at onset. *British Journal of Psychiatry*, *202*(s54), s5–s10. https://doi.org/10.1192/bjp.bp.112.119164

Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

Kaplan, R. M., & Saccuzzo, D. P. (2001). *Psychological testing: Principles, applications, and issues, 5th ed* (pp. xxiii, 708). Wadsworth/Thomson Learning.

KATSCHNIG, H. (2010). Are psychiatrists an endangered species? Observations on internal and external challenges to the profession. *World Psychiatry*, *9*(1), 21–28.

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., Williamson, D., & Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): Initial Reliability and Validity Data. *Journal of the American Academy of Child & Adolescent Psychiatry*, *36*(7), 980–988. https://doi.org/10.1097/00004583-199707000-00021

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3149–3157.

Kessler, R. C. (2007). Psychiatric epidemiology: Challenges and opportunities. *International Review of Psychiatry*, *19*(5), 509–521. https://doi.org/10.1080/09540260701564914

Kessler, R. C. (2011). The National Comorbidity Survey (NCS) and its Extensions. In *Textbook of Psychiatric Epidemiology* (pp. 221–241). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470976739.ch14

KESSLER, R. C., ANGERMEYER, M., ANTHONY, J. C., DE GRAAF, R., DEMYTTENAERE, K., GASQUET, I., DE GIROLAMO, G., GLUZMAN, S., GUREJE, O., HARO, J. M., KAWAKAMI, N., KARAM, A., LEVINSON, D., MEDINA MORA, M. E., OAKLEY BROWNE, M. A., POSADA-VILLA, J., STEIN, D. J., ADLEY TSANG, C. H., AGUILAR-GAXIOLA, S., … ÜSTÜN, T. B. (2007). Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry*, *6*(3), 168–176.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders

in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 593–602. https://doi.org/10.1001/archpsyc.62.6.593

Kondo, T. (2015). How and Why is Autism Spectrum Disorder Misdiagnosed in Adult Patients? - From Diagnostic Problem to Management for Adjustment -. *Mental Health in Family Medicine*, *11*, 73–88. https://doi.org/10.25149/1756-8358.1102011

Konishcheva, K. (2023). *Diagnosis-predictor* [Python]. https://github.com/charlie42/diagnosis-predictor (Original work published 2022)

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kooij, J. J. S. (2012). *Adult ADHD: Diagnostic Assessment and Treatment*. Springer Science & Business Media.

Kramer, E., Koo, B., Restrepo, A., Koyama, M., Neuhaus, R., Pugh, K., Andreotti, C., & Milham, M. (2020). Diagnostic Associations of Processing Speed in a Transdiagnostic, Pediatric Sample. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-66892-z

Krishnan, V., & Krishnan, V. (n.d.). *The Early Child Development Instrument (EDI): An item analysis using Classical Test Theory (CTT) on Alberta's data*.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kronenberger, W. G., & Dunn, D. W. (2003). Learning disorders. *Neurologic Clinics*, *21*(4), 941–952. https://doi.org/10.1016/S0733-8619(03)00010-0

Kuhle, S., Maguire, B., Zhang, H., Hamilton, D., Allen, A. C., Joseph, K. S., & Allen, V. M. (2018). Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: A retrospective cohort study. *BMC Pregnancy and Childbirth*, *18*(1), 333. https://doi.org/10.1186/s12884-018-1971-2

Labaye-Prévot, N., Weens, B., & Moltrecht, B. (2022). *Vers une école promotrice de santé: Guide du diagnostic à l'action*. Presses de l'École des hautes études en santé publique.

Lam, K. S. L., & Aman, M. G. (2007). The Repetitive Behavior Scale-Revised: Independent validation in individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *37*(5), 855–866. https://doi.org/10.1007/s10803-006-0213-z

Le label Édusanté. (n.d.). Ministère de l'Éducation nationale. https://eduscol.education.fr/document/33860/download?attachment

Le livret de parcours inclusif (LPI). (2023, September). Eduscol. https://eduscol.education.fr/2506/le-livret-de-parcours-inclusif-lpi

Leathart, T., Frank, E., Holmes, G., & Pfahringer, B. (2017). Probability Calibration Trees. *Proceedings of the Ninth Asian Conference on Machine Learning*, 145–160. https://proceedings.mlr.press/v77/leathart17a.html

Leeman, E. (1999). Limitations of epidemiological field data for mental health policy decisions. *International Journal of Psychiatry in Clinical Practice*, *3*(3), 155–157. https://doi.org/10.3109/13651509909022728

Les établissements régionaux d'enseignement adapté. (2023, October). éduscol | Ministère de l'Éducation nationale et de la Jeunesse - Direction générale de l'enseignement scolaire. https://eduscol.education.fr/1178/les-etablissements-regionaux-d-enseignement-adapte

Les réseaux d'aides spécialisées aux élèves en difficulté (Rased). (2014, août). Ministère de l'Education Nationale et de la Jeunesse. https://www.education.gouv.fr/les-reseaux-d-aides-specialisees-aux-eleves-en-difficulte-rased-11312

Lewis, F. I., & Torgerson, P. R. (2012). A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic. *Emerging Themes in Epidemiology*, *9*(1), 9. https://doi.org/10.1186/1742-7622-9-9

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature Selection: A Data Perspective. *ACM Computing Surveys*, *50*(6), 1–45. https://doi.org/10.1145/3136625

Liao, Y., Ameyaw, M. A., Liang, C., & Li, W. (2023). Research on the Effect of Evidence-Based Intervention on Improving Students' Mental Health Literacy Led by Ordinary Teachers: A Meta-Analysis. *International Journal of Environmental Research and Public Health*, *20*(2), Article 2. https://doi.org/10.3390/ijerph20020949

Liddle, E. B., Batty, M. J., & Goodman, R. (2009). The Social Aptitudes Scale: An initial validation. *Social Psychiatry and Psychiatric Epidemiology*, *44*(6), 508–513. https://doi.org/10.1007/s00127-008-0456-4

Liu, N. H., Daumit, G. L., Dua, T., Aquila, R., Charlson, F., Cuijpers, P., Druss, B., Dudek, K., Freeman, M., Fujii, C., Gaebel, W., Hegerl, U., Levav, I., Munk Laursen, T., Ma, H., Maj, M., Elena Medina-Mora, M., Nordentoft, M., Prabhakaran, D., … Saxena, S. (2017). Excess mortality in persons with severe mental disorders: A multilevel intervention framework and priorities for clinical practice, policy and research agendas. *World Psychiatry*, *16*(1), 30–40. https://doi.org/10.1002/wps.20384

Liu, Y., Hankey, J., Cao, B., & Chokka, P. (2021). Screening for major depressive disorder in a tertiary mental health centre using EarlyDetect: A machine learning-based pilot study. *Journal of Affective Disorders Reports*, *3*, 100062. https://doi.org/10.1016/j.jadr.2020.100062

Liu, Y. S., Cao, B., & Chokka, P. R. (2023). Screening for Adulthood ADHD and Comorbidities in a Tertiary Mental Health Center Using EarlyDetect: A Machine Learning-Based Pilot Study. *Journal of Attention Disorders*, *27*(3), 324–331. https://doi.org/10.1177/10870547221136228

Liu, Y. S., Chokka, S., Cao, B., & Chokka, P. R. (2021). Screening for bipolar disorder in a tertiary mental health centre using EarlyDetect: A machine learning-based pilot study. *Journal of Affective Disorders Reports*, *6*, 100215. https://doi.org/10.1016/j.jadr.2021.100215

Loney, J. (1982). Hyperactivity, inattention and aggression in clinical practice. *Advances in Behavioral Pediatrics*, *2*, 113–147.

Lu, F., & Petkova, E. (2014). A comparative study of variable selection methods in the context of developing psychiatric screening instruments. *Statistics in Medicine*, *33*(3), 401–421. https://doi.org/10.1002/sim.5937

Lubke, G. H., Hudziak, J. J., Derks, E. M., van Bijsterveldt, T. C. E. M., & Boomsma, D. I. (2009). Maternal ratings of attention problems in ADHD: Evidence for the existence of a continuum. *Journal of the American Academy of Child and Adolescent Psychiatry*, *48*(11), 1085–1093. https://doi.org/10.1097/CHI.0b013e3181ba3dbb

MacDonald, K., Fainman-Adelman, N., Anderson, K. K., & Iyer, S. N. (2018). Pathways to mental health services for young people: A systematic review. *Social Psychiatry*

and Psychiatric Epidemiology, *53*(10), 1005–1038. https://doi.org/10.1007/s00127-018-1578-y

MacDonald, K., Ferrari, M., Fainman-Adelman, N., & Iyer, S. N. (2021). Experiences of pathways to mental health services for young people and their carers: A qualitative meta-synthesis review. *Social Psychiatry and Psychiatric Epidemiology*, *56*(3), 339–361. https://doi.org/10.1007/s00127-020-01976-9

Magno, C. (2009). *Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data* (SSRN Scholarly Paper 1426043). https://papers.ssrn.com/abstract=1426043

Mandalia, D., Ford, T., Hill, S., Sadler, K., Vizard, T., Goodman, A., Goodman, R., & McManus, S. (2018). *Mental Health of Children and Young People in England 2021*. NHS. https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2021-follow-up-to-the-2017-survey

Margolis, A. E., Broitman, J., Davis, J. M., Alexander, L., Hamilton, A., Liao, Z., Banker, S., Thomas, L., Ramphal, B., Salum, G. A., Merikangas, K., Goldsmith, J., Paus, T., Keyes, K., & Milham, M. P. (2020). Estimated Prevalence of Nonverbal Learning Disability Among North American Children and Adolescents. *JAMA Network Open*, *3*(4), e202551. https://doi.org/10.1001/jamanetworkopen.2020.2551

McGorry, P. D., Goldstone, S. D., Parker, A. G., Rickwood, D. J., & Hickie, I. B. (2014a). Cultures for mental health care of young people: An Australian blueprint for reform. *The Lancet Psychiatry*, *1*(7), 559–568. https://doi.org/10.1016/S2215-0366(14)00082-0

McGorry, P. D., Goldstone, S. D., Parker, A. G., Rickwood, D. J., & Hickie, I. B. (2014b). Cultures for mental health care of young people: An Australian blueprint for reform. *The Lancet Psychiatry*, *1*(7), 559–568. https://doi.org/10.1016/S2215-0366(14)00082-0

Merikangas, K. R., He, J., Rapoport, J., Vitiello, B., & Olfson, M. (2013). Medication Use in US Youth With Mental Disorders. *JAMA Pediatrics*, *167*(2), 141–148. https://doi.org/10.1001/jamapediatrics.2013.431

Merikangas, K. R., & Kalaydjian, A. (2007). Magnitude and impact of comorbidity of mental disorders from epidemiologic surveys. *Current Opinion in Psychiatry*, *20*(4), 353–358. https://doi.org/10.1097/YCO.0b013e3281c61dc5

Merikangas, K. R., Nakamura, E. F., & Kessler, R. C. (2009). Epidemiology of mental disorders in children and adolescents. *Dialogues in Clinical Neuroscience*, *11*(1), 7–20. https://doi.org/10.31887/DCNS.2009.11.1/krmerikangas

Merry, S. N., Hetrick, S. E., Cox, G. R., Brudevold-Iversen, T., Bir, J. J., & McDowell, H. (2011). Psychological and educational interventions for preventing depression in

children and adolescents. *The Cochrane Database of Systematic Reviews*, *12*, CD003380. https://doi.org/10.1002/14651858.CD003380.pub3

Mettre en confiance l'élève pour faciliter sa production orale. (n.d.). Retrieved November 15, 2023, from https://www.reseau-canope.fr/cap-ecole-inclusive/amenager-et-adapter/fiche-adaptation/mettre-en-confiance-leleve-pour-faciliter-sa-production-orale.html

mhdb: Mental health database. (2021). [Computer software]. Child Mind Institute. https://github.com/ChildMindInstitute/mhdb (Original work published 2020)

Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science*, *91*, 919–926. https://doi.org/10.1016/j.procs.2016.07.111

Milgrom, J., & Gemmill, A. W. (2014). Screening for perinatal depression. *Best Practice & Research Clinical Obstetrics & Gynaecology*, *28*(1), 13–23. https://doi.org/10.1016/j.bpobgyn.2013.08.014

Mise en place du parcours éducatif de santé pour tous les élèves. (2016, January 28). Ministère de l'Education Nationale et de la Jeunesse. https://www.education.gouv.fr/bo/16/Hebdo5/MENE1601852C.htm

Mitropoulos, G. B. (2018). The DSM-ICD diagnostic approach as an essential bridge between the patient and the "big data." *Psychiatriki*, *29*(3), 249–256. https://doi.org/10.22365/jpsych.2018.293.249

Molnar, C. (2023). *Chapter 5 Interpretable Models | Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/simple.html

Moore, S. A., Widales-Benitez, O., Carnazzo, K. W., Kim, E. K., Moffa, K., & Dowdy, E. (2015). Conducting Universal Complete Mental Health Screening via Student Self-Report. *Contemporary School Psychology*, *19*(4), 253–267. https://doi.org/10.1007/s40688-015-0062-x

Moro, M.-R., & Brison, J.-L. (2016). *Mission bien-être et santé des jeunes*.

Mugnaini, D., Lassi, S., La Malfa, G., & Albertini, G. (2009). Internalizing correlates of dyslexia. *World Journal of Pediatrics*, *5*(4), 255–264. https://doi.org/10.1007/s12519-009-0049-7

Mujahid, M. S., Diez Roux, A. V., Morenoff, J. D., & Raghunathan, T. (2007). Assessing the measurement properties of neighborhood scales: From psychometrics to ecometrics. *American Journal of Epidemiology*, *165*(8), 858–867. https://doi.org/10.1093/aje/kwm040

Muris, P., Merckelbach, H., Kindt, M., Bögels, S., Dreessen, L., Dorp, C. V., Habets, A., Rosmuller, S., & Snieder, N. (2001). The utility of screen for child anxiety related emotional disorders (scared) as a tool for identifying children at high risk for prevalent anxiety disorders. *Anxiety, Stress & Coping*, *14*(3), 265–283. https://doi.org/10.1080/10615800108248357

Murphy, J. M., Reede, J., Jellinek, M. S., & Bishop, S. J. (1992). Screening for Psychosocial Dysfunction in Inner-City Children: Further Validation of the Pediatric Symptom Checklist. *Journal of the American Academy of Child & Adolescent Psychiatry*, *31*(6), 1105–1111. https://doi.org/10.1097/00004583-199211000-00019

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, *4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

Natarajan, P., Frenzel, J. C., & Smaltz, D. H. (2017). *Demystifying Big Data and Machine Learning for Healthcare* (1st edition). CRC Press.

Nielsen, L. G., Rimvall, M. K., Clemmensen, L., Munkholm, A., Elberling, H., Olsen, E. M., Rask, C. U., Skovgaard, A. M., & Jeppesen, P. (2019). The predictive validity of the Strengths and Difficulties Questionnaire in preschool age to identify mental disorders in preadolescence. *PLOS ONE*, *14*(6), e0217707. https://doi.org/10.1371/journal.pone.0217707

Nooner, K. B., Colcombe, S. J., Tobe, R. H., Mennes, M., Benedict, M. M., Moreno, A. L., Panek, L. J., Brown, S., Zavitz, S. T., Li, Q., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R. T., Kamiel, S. M., Anwar, A. R., Hinz, C. M., Kaplan, M. S., Rachlin, A. B., … Milham, M. P. (2012). The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Frontiers in Neuroscience*, *6*, 152. https://doi.org/10.3389/fnins.2012.00152

Nusinovici, S., Tham, Y. C., Chak Yan, M. Y., Wei Ting, D. S., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

O'Connor, C. A., Dyson, J., Cowdell, F., & Watson, R. (2018). Do universal school-based mental health promotion programmes improve the mental health and emotional wellbeing of young people? A literature review. *Journal of Clinical Nursing*, *27*(3–4), e412–e426. https://doi.org/10.1111/jocn.14078

O'Farrell, P., Wilson, C., & Shiel, G. (2023). Teachers' perceptions of the barriers to assessment of mental health in schools with implications for educational policy: A systematic review. *The British Journal of Educational Psychology*, *93*(1), 262–282. https://doi.org/10.1111/bjep.12553

Olariu, E., Forero, C. G., Castro-Rodriguez, J. I., Rodrigo-Calvo, M. T., Álvarez, P., Martín-López, L. M., Sánchez-Toto, A., Adroher, N. D., Blasco-Cubedo, M. J., Vilagut, G., Fullana, M. A., & Alonso, J. (2015). Detection of Anxiety Disorders in Primary Care: A Meta-Analysis of Assisted and Unassisted Diagnoses. *Depression and Anxiety*, *32*(7), 471–484. https://doi.org/10.1002/da.22360

Oommen, T., Baise, L. G., & Vogel, R. M. (2011). Sampling Bias and Class Imbalance in Maximum-likelihood Logistic Regression. *Mathematical Geosciences*, *43*(1), 99–120. https://doi.org/10.1007/s11004-010-9311-8

Ormel, J., Oerlemans, A. M., Raven, D., Laceulle, O. M., Hartman, C. A., Veenstra, R., Verhulst, F. C., Vollebergh, W., Rosmalen, J. G. M., Reijneveld, S. A., & Oldehinkel, A. J. (2017). Functional outcomes of child and adolescent mental disorders. Current disorder most important but psychiatric history matters as well. *Psychological Medicine*, *47*(7), 1271–1282. https://doi.org/10.1017/S0033291716003445

Ormel, J., Petukhova, M., Chatterji, S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., Bromet, E. J., Burger, H., Demyttenaere, K., de Girolamo, G., Haro, J. M., Hwang, I., Karam, E., Kawakami, N., Lepine, J. P., Medina-Mora, M. E., Posada-Villa, J., Sampson, N., Scott, K., … Kessler, R. C. (2008). Disability and treatment of specific mental and physical disorders across the world: Results from the WHO World Mental Health Surveys. *The British Journal of Psychiatry : The Journal of Mental Science*, *192*(5), 368–375. https://doi.org/10.1192/bjp.bp.107.039107

Ozernov-Palchik, O., Norton, E. S., Sideridis, G., Beach, S. D., Wolf, M., Gabrieli, J. D. E., & Gaab, N. (2017). Longitudinal stability of pre-reading skill profiles of kindergarten children: Implications for early screening and theories of reading. *Developmental Science*, *20*(5), 10.1111/desc.12471. https://doi.org/10.1111/desc.12471

Panesar, A. (2019). *Machine Learning and AI for Healthcare* (1st ed. edition). Apress.

Parnas, J. (2015). Differential diagnosis and current polythetic classification. *World Psychiatry*, *14*(3), 284–287. https://doi.org/10.1002/wps.20239

Patalay, P., Gondek, D., Moltrecht, B., Giese, L., Curtin, C., Stanković, M., & Savka, N. (2017). Mental health provision in schools: Approaches and interventions in 10 European countries. *Global Mental Health*, *4*, e10. https://doi.org/10.1017/gmh.2017.6

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Petterson, B., Bourke, J., Leonard, H., Jacoby, P., & Bower, C. (2007). Co-occurrence of birth defects and intellectual disability. *Paediatric and Perinatal Epidemiology*, *21*(1), 65–75. https://doi.org/10.1111/j.1365-3016.2007.00774.x

Polanczyk, G. V., Salum, G. A., Sugaya, L. S., Caye, A., & Rohde, L. A. (2015). Annual Research Review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *Journal of Child Psychology and Psychiatry*, *56*(3), 345–365. https://doi.org/10.1111/jcpp.12381

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, *15*(11), 1119–1125. https://doi.org/10.1016/0167-8655(94)90127-9

Qui sommes-nous. (n.d.). Réseau Canopé. Retrieved November 14, 2023, from https://www.reseau-canope.fr/qui-sommes-nous.html

Ranganathan, P., & Aggarwal, R. (2018). Common pitfalls in statistical analysis: Understanding the properties of diagnostic tests – Part 1. *Perspectives in Clinical Research*, *9*(1), 40–43. https://doi.org/10.4103/picr.PICR_170_17

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software*, *3*(24), 638. https://doi.org/10.21105/joss.00638

Raschka, S. (2020). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning* (arXiv:1811.12808). arXiv. http://arxiv.org/abs/1811.12808

Regier, D. A., Kaelber, C. T., Rae, D. S., Farmer, M. E., Knauper, B., Kessler, R. C., & Norquist, G. S. (1998). Limitations of Diagnostic Criteria and Assessment Instruments for Mental Disorders: Implications for Research and Policy. *Archives of General Psychiatry*, *55*(2), 109–115. https://doi.org/10.1001/archpsyc.55.2.109

Regier, D. A., Kuhl, E. A., & Kupfer, D. J. (2013). The DSM-5: Classification and criteria changes. *World Psychiatry*, *12*(2), 92–98. https://doi.org/10.1002/wps.20050

Reise, S. P., & Rodriguez, A. (2016). Item response theory and the measurement of psychiatric constructs: Some empirical and conceptual issues and challenges.

Psychological Medicine, *46*(10), 2025–2039. https://doi.org/10.1017/S0033291716000520

Reise, S. P., & Waller, N. G. (2009). Item Response Theory and Clinical Measurement. *Annual Review of Clinical Psychology*, *5*(1), 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Répertoire des interventions efficaces ou prometteuses en prévention et promotion de la santé. (n.d.). Retrieved November 14, 2023, from https://www.santepubliquefrance.fr/a-propos/services/interventions-probantes-ou-prometteuses-en-prevention-et-promotion-de-la-sante/repertoire-des-interventions-efficaces-ou-prometteuses-en-prevention-et-promotion-de-la-sante

Rescorla, L. A., Bochicchio, L., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I., Bilenberg, N., Bird, H., Dobrean, A., Erol, N., Fombonne, E., Fonseca, A., Frigerio, A., Fung, D. S. S., Lambert, M. C., Leung, P. W. L., Liu, X., Marković, I., Markovic, J., … Verhulst, F. C. (2014). Parent–Teacher Agreement on Children's Problems in 21 Societies. *Journal of Clinical Child & Adolescent Psychology*, *43*(4), 627–642. https://doi.org/10.1080/15374416.2014.900719

Revelle, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research* (2.3.9) [Computer software]. https://cran.r-project.org/web/packages/psych/index.html

Reyes, A. D. L., & Langer, D. A. (2018). Assessment and the Journal of Clinical Child and Adolescent Psychology's Evidence Base Updates Series: Evaluating the Tools for Gathering Evidence. *Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *47*(3), 357–365. https://doi.org/10.1080/15374416.2018.1458314

Richardson, L. P., McCauley, E., Grossman, D. C., McCarty, C. A., Richards, J., Russo, J. E., Rockhill, C., & Katon, W. (2010). Evaluation of the Patient Health Questionnaire (PHQ-9) for Detecting Major Depression among Adolescents. *Pediatrics*, *126*(6), 1117–1123. https://doi.org/10.1542/peds.2010-0852

Richardson, W. S., & Wilson, M. C. (2015). The Process of Diagnosis. In G. Guyatt, D. Rennie, M. O. Meade, & D. J. Cook (Eds.), *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 3rd ed* (1–Book, Section). McGraw-Hill Education. jamaevidence.mhmedical.com/content.aspx?aid=1183877015

Ricky, C., Siobhan, O., Nawaf, M., & Elliot M., G. (2017). Factors associated with delayed diagnosis of mood and/or anxiety disorders. *Health Promotion and Chronic Disease Prevention in Canada : Research, Policy and Practice*, *37*(5), 137–148.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26.1*, 108–116. https://doi.org/10.7334/psicothema2013.260

Risal, A. (2011). Common Mental Disorders. *Kathmandu University Medical Journal*, *9*(3), Article 3. https://doi.org/10.3126/kumj.v9i3.6308

Roefs, A., Fried, E. I., Kindt, M., Martijn, C., Elzinga, B., Evers, A. W. M., Wiers, R. W., Borsboom, D., & Jansen, A. (2022). A new science of mental disorders: Using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology. *Behaviour Research and Therapy*, *153*, 104096. https://doi.org/10.1016/j.brat.2022.104096

Roeser, R. W., & Midgley, C. (1997). Teachers' Views of Issues Involving Students' Mental Health. *The Elementary School Journal*, *98*(2), 115–133. https://doi.org/10.1086/461887

Russell, P. S. S., Nair, M. K. C., Russell, S., Subramaniam, V. S., Sequeira, A. Z., Nazeema, S., & George, B. (2013). ADad 2: The Validation of the Screen for Child Anxiety Related Emotional Disorders for Anxiety Disorders Among Adolescents in a Rural Community Population in India. *The Indian Journal of Pediatrics*, *80*(2), 139–143. https://doi.org/10.1007/s12098-013-1233-2

RUTTER, M., BAILEY, A., & LORD, C. (2003). *(SCQ) Social Communication Questionnaire*. https://www.wpspublish.com/scq-social-communication-questionnaire.html

Salaheddin, K., & Mason, B. (2016). Identifying barriers to mental health help-seeking among young adults in the UK: A cross-sectional survey. *The British Journal of General Practice*, *66*(651), e686–e692. https://doi.org/10.3399/bjgp16X687313

Sanfilippo, J., Ness, M., Petscher, Y., Rappaport, L., Zuckerman, B., & Gaab, N. (2020). Reintroducing Dyslexia: Early Identification and Implications for Pediatric Practice. *Pediatrics*, *146*(1), e20193046. https://doi.org/10.1542/peds.2019-3046

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Savage, H., Murray, J., Hatch, S. L., Hotopf, M., Evans-Lacko, S., & Brown, J. S. L. (2016). Exploring Professional Help-Seeking for Mental Disorders. *Qualitative Health Research*, *26*(12), 1662–1673. https://doi.org/10.1177/1049732315591483

Sayal, K., Taylor, E., Beecham, J., & Byrne, P. (2002). Pathways to care in children at risk of attention-deficit hyperactivity disorder. *British Journal of Psychiatry*, *181*(1), 43–48. https://doi.org/10.1192/bjp.181.1.43

Schelbe, L., Pryce, J., Petscher, Y., Fien, H., Stanley, C., Gearin, B., & Gaab, N. (2022). Dyslexia in the Context of Social Work: Screening and Early Intervention. *Families in Society*, *103*(3), 269–280. https://doi.org/10.1177/10443894211042323

Schomerus, G., Stolzenburg, S., Freitag, S., Speerforck, S., Janowitz, D., Evans-Lacko, S., Muehlan, H., & Schmidt, S. (2019). Stigma as a barrier to recognizing personal mental illness and seeking help: A prospective study among untreated persons with mental illness. *European Archives of Psychiatry and Clinical Neuroscience*, *269*(4), 469–479. https://doi.org/10.1007/s00406-018-0896-0

Schools4Health—Schools for Health. (n.d.). Retrieved November 15, 2023, from https://schools4health.eu/project

Schultebraucks, K., Stevens, J. S., Michopoulos, V., Maples-Keller, J., Lyu, J., Smith, R. N., Rothbaum, B. O., Ressler, K. J., Galatzer-Levy, I. R., & Powers, A. (2023). Development and validation of a brief screener for posttraumatic stress disorder risk in emergency medical settings. *General Hospital Psychiatry*, *81*, 46–50. https://doi.org/10.1016/j.genhosppsych.2023.01.012

Setia, M. S. (2016). Methodology Series Module 3: Cross-sectional Studies. *Indian Journal of Dermatology*, *61*(3), 261–264. https://doi.org/10.4103/0019-5154.182410

Sharp, C., Croudace, T. J., Goodyer, I. M., & Amtmann, D. (2005). The Strength and Difficulties Questionnaire: Predictive validity of parent and teacher ratings for help-seeking behaviour over one year. *Educational and Child Psychology*, *22*(3), 28–44. https://doi.org/10.53841/bpsecp.2005.22.3.28

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(09), 1426–1448. https://doi.org/10.1017/S0033291719000151

Shelemy, L., Harvey, K., & Waite, P. (2019). Supporting students' mental health in schools: What do teachers want and need? *Emotional and Behavioural Difficulties*, *24*(1), 100–116. https://doi.org/10.1080/13632752.2019.1582742

Siceloff, E. R., Bradley, W. J., & Flory, K. (2017). Universal Behavioral/Emotional Health Screening in Schools: Overview and Feasibility. *Report on Emotional & Behavioral Disorders in Youth*, *17*(2), 32–38.

Silva, I. dos S. (1999). *Cancer Epidemiology: Principles and Methods*. IARC.

Simms, L. J. (2008). Classical and Modern Methods of Psychological Scale Construction. *Social and Personality Psychology Compass*, *2*(1), 414–433. https://doi.org/10.1111/j.1751-9004.2007.00044.x

Simons, J. S., & Gaher, R. M. (2005). The Distress Tolerance Scale: Development and Validation of a Self-Report Measure. *Motivation and Emotion*, *29*(2), 83–102. https://doi.org/10.1007/s11031-005-7955-3

S'informer. (n.d.). Retrieved November 14, 2023, from https://www.reseau-canope.fr/cap-ecole-inclusive/sinformer.html

Sklearn.model_selection.LeaveOneOut. (n.d.). Scikit-Learn. Retrieved November 14, 2023, from https://scikit-learn/stable/modules/generated/sklearn.model_selection.LeaveOneOut.html

Soares, N., Evans, T., & Patel, D. R. (2018). Specific learning disability in mathematics: A comprehensive review. *Translational Pediatrics*, *7*(1), 48–62. https://doi.org/10.21037/tp.2017.08.03

Sokal, L., & Sharma, U. (2013). Canadian In-service Teachers' Concerns, Efficacy, and Attitudes about Inclusive Teaching. *Exceptionality Education International*, *23*, 59–71. https://doi.org/10.5206/eei.v23i1.7704

Solmi, M., Radua, J., Olivola, M., Croce, E., Soardo, L., Salazar de Pablo, G., Il Shin, J., Kirkbride, J. B., Jones, P., Kim, J. H., Kim, J. Y., Carvalho, A. F., Seeman, M. V., Correll, C. U., & Fusar-Poli, P. (2022). Age at onset of mental disorders worldwide: Large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry*, *27*(1), Article 1. https://doi.org/10.1038/s41380-021-01161-7

SPF. (2022, février). *Les compétences psychosociales: Un référentiel pour un déploiement auprès des enfants et des jeunes. Synthèse de l'état des connaissances scientifiques et théoriques réalisé en 2021*. https://www.santepubliquefrance.fr/import/les-competences-psychosociales-un-referentiel-pour-un-deploiement-aupres-des-enfants-et-des-jeunes.-synthese-de-l-etat-des-connaissances-scientif

Spiess, A.-N., & Neumeyer, N. (2010). An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: A Monte Carlo approach. *BMC Pharmacology*, *10*, 6. https://doi.org/10.1186/1471-2210-10-6

Stigma, Prejudice and Discrimination Against People with Mental Illness. (n.d.). Retrieved November 14, 2023, from https://www.psychiatry.org:443/patients-families/stigma-and-discrimination

Stringaris, A., Goodman, R., Ferdinando, S., Razdan, V., Muhrer, E., Leibenluft, E., & Brotman, M. A. (2012). The Affective Reactivity Index: A concise irritability scale for clinical and research settings. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *53*(11), 1109–1117. https://doi.org/10.1111/j.1469-7610.2012.02561.x

Su, L., Wang, K., Fan, F., Su, Y., & Gao, X. (2008). Reliability and validity of the screen for child anxiety related emotional disorders (SCARED) in Chinese children. *Journal of Anxiety Disorders*, *22*(4), 612–621. https://doi.org/10.1016/j.janxdis.2007.05.011

Surajustement. (n.d.). Retrieved November 15, 2023, from https://fr.mathworks.com/discovery/overfitting.html

SurveyMonkey. (n.d.). SurveyMonkey. Retrieved November 16, 2023, from https://www.surveymonkey.com/

Sveen, T. H., Berg-Nielsen, T. S., Lydersen, S., & Wichstrøm, L. (2013). Detecting Psychiatric Disorders in Preschoolers: Screening With the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(7), 728–736. https://doi.org/10.1016/j.jaac.2013.04.010

Swanson, J. M., Kraemer, H. C., Hinshaw, S. P., Arnold, L. E., Conners, C. K., Abikoff, H. B., Clevenger, W., Davies, M., Elliott, G. R., Greenhill, L. L., Hechtman, L., Hoza, B., Jensen, P. S., March, J. S., Newcorn, J. H., Owens, E. B., Pelham, W. E., Schiller, E., Severe, J. B., … Wu, M. (2001). Clinical relevance of the primary findings of the MTA: Success rates based on severity of ADHD and ODD symptoms at the end of treatment. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*(2), 168–179. https://doi.org/10.1097/00004583-200102000-00011

Swanson, J., Posner, M., Fusella, J., Wasdell, M., Sommer, T., & Fan, J. (2001). Genes and attention deficit hyperactivity disorder. *Current Psychiatry Reports*, *3*(2), 92–100. https://doi.org/10.1007/s11920-001-0005-2

Symptom Checker. (n.d.). Child Mind Institute. Retrieved November 14, 2023, from https://childmind.org/symptomchecker/

Tartarisco, G., Cicceri, G., Di Pietro, D., Leonardi, E., Aiello, S., Marino, F., Chiarotti, F., Gagliano, A., Arduino, G. M., Apicella, F., Muratori, F., Bruneo, D., Allison, C., Cohen, S. B., Vagni, D., Pioggia, G., & Ruta, L. (2021). Use of Machine Learning to Investigate the Quantitative Checklist for Autism in Toddlers (Q-CHAT) towards

Early Autism Screening. *Diagnostics (Basel, Switzerland)*, *11*(3), 574. https://doi.org/10.3390/diagnostics11030574

Tesfaw, G., Kibru, B., & Ayano, G. (2020). Prevalence and factors associated with higher levels of perceived stigma among people with schizophrenia Addis Ababa, Ethiopia. *International Journal of Mental Health Systems*, *14*(1), 19. https://doi.org/10.1186/s13033-020-00348-9

The Stavros Niarchos Foundation Global Center for Child and Adolescent Mental Health. (n.d.). Child Mind Institute. Retrieved November 15, 2023, from https://childmind.org/global/snf-global-center/

The World Mental Health Survey Initiative. (n.d.). Retrieved November 14, 2023, from https://www.hcp.med.harvard.edu/wmh/

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tunks, A., Berry, C., Strauss, C., Nyikavaranda, P., & Ford, E. (2023). Patients' perspectives of barriers and facilitators to accessing support through primary care for common mental health problems in England: A systematic review. *Journal of Affective Disorders*, *338*, 329–340. https://doi.org/10.1016/j.jad.2023.06.035

Tutun, S., Johnson, M. E., Ahmed, A., Albizri, A., Irgil, S., Yesilkaya, I., Ucar, E. N., Sengun, T., & Harfouche, A. (2023). An AI-based Decision Support System for Predicting Mental Health Disorders. *Information Systems Frontiers*, *25*(3), 1261–1276. https://doi.org/10.1007/s10796-022-10282-5

Upright. (n.d.). *UPRIGHT*. Retrieved November 15, 2023, from https://uprightproject.eu/

van den Goorbergh, R., van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, *29*(9), 1525–1534. https://doi.org/10.1093/jamia/ocac093

Vaudreuil, C. A. H., Faraone, S. V., Di Salvo, M., Wozniak, J. R., Wolenski, R. A., Carrellas, N. W., & Biederman, J. (2019). The morbidity of subthreshold pediatric bipolar disorder: A systematic literature review and meta-analysis. *Bipolar Disorders*, *21*(1), 16–27. https://doi.org/10.1111/bdi.12734

Vaughn, S., Wanzek, J., Woodruff, T., & Linan-Thompson, S. (2007). *Prevention and early identification of students with reading disabilities.*

Venkatesh, B., & Anuradha, J. (2019). A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*, *19*(1), 3–26. https://doi.org/10.2478/cait-2019-0001

Vermani, M., Marcus, M., & Katzman, M. A. (2011). Rates of Detection of Mood and Anxiety Disorders in Primary Care: A Descriptive, Cross-Sectional Study. *The Primary Care Companion for CNS Disorders*, *13*(2), 27211. https://doi.org/10.4088/PCC.10m01013

Vessonen, E. (2020). The Complementarity of Psychometrics and the Representational Theory of Measurement. *The British Journal for the Philosophy of Science*, *71*(2), 415–442. https://doi.org/10.1093/bjps/axy032

Viaene, S., Derrig, R. A., & Dedene, G. (2004). A case study of applying boosting naive bayes to claim fraud diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, *16*(5), 612–620. https://doi.org/10.1109/TKDE.2004.1277822

WANG, P. S., ANGERMEYER, M., BORGES, G., BRUFFAERTS, R., TAT CHIU, W., DE GIROLAMO, G., FAYYAD, J., GUREJE, O., HARO, J. M., HUANG, Y., KESSLER, R. C., KOVESS, V., LEVINSON, D., NAKANE, Y., OAKLEY BROWN, M. A., ORMEL, J. H., POSADA-VILLA, J., AGUILAR-GAXIOLA, S., ALONSO, J., … ÜSTÜN, T. B. (2007). Delay and failure in treatment seeking after first onset of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry*, *6*(3), 177–185.

Ward, M. M. (2013). ESTIMATING DISEASE PREVALENCE AND INCIDENCE USING ADMINISTRATIVE DATA: SOME ASSEMBLY REQUIRED. *The Journal of Rheumatology*, *40*(8), 1241–1243. https://doi.org/10.3899/jrheum.130675

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children | Fifth Edition*. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Cognition-%26-Neuro/Gifted-%26-Talented/Wechsler-Intelligence-Scale-for-Children-%7C-Fifth-Edition-/p/100000771.html

What is a Segpa class? (2023, August 17). Service-Public.Fr. https://www.service-public.fr/particuliers/vosdroits/F32752?lang=en

What is Prevalence? (n.d.). NIMH. Retrieved November 14, 2023, from https://www.nimh.nih.gov/health/statistics/what-is-prevalence

Willcutt, E. G., Boada, R., Riddle, M. W., Chhabildas, N., DeFries, J. C., & Pennington, B. F. (2011). Colorado Learning Difficulties Questionnaire:Validation of a parent-report screening measure. *Psychological Assessment*, *23*(3), 778–791. https://doi.org/10.1037/a0023290

Wittchen, H.-U., Mühlig, S., & Beesdo, K. (2003). Mental disorders in primary care. *Dialogues in Clinical Neuroscience*, *5*(2), 115. https://doi.org/10.31887/DCNS.2003.5.2/huwittchen

Wolraich, M. L., Bard, D. E., Neas, B., Doffing, M., & Beck, L. (2013). The Psychometric Properties of the Vanderbilt Attention-Deficit Hyperactivity Disorder Diagnostic Teacher Rating Scale in a Community Population. *Journal of Developmental & Behavioral Pediatrics*, *34*(2), 83. https://doi.org/10.1097/DBP.0b013e31827d55c3

Wolraich, M. L., Feurer, I. D., Hannah, J. N., Baumgaertel, A., & Pinnock, T. Y. (1998). Obtaining systematic teacher reports of disruptive behavior disorders utilizing DSM-IV. *Journal of Abnormal Child Psychology*, *26*(2), 141–152. https://doi.org/10.1023/a:1022673906401

Wood, B. J., & Ellis, F. (2022). Universal Mental Health Screening Practices in Midwestern Schools: A Window of Opportunity for School Psychologist Leadership and Role Expansion? *Contemporary School Psychology*. https://doi.org/10.1007/s40688-022-00430-8

Wood, B. J., & McDaniel, T. (2020). A preliminary investigation of universal mental health screening practices in schools. *Children and Youth Services Review*, *112*, 104943. https://doi.org/10.1016/j.childyouth.2020.104943

World Health Organization. (2019). *International statistical classification of diseases and related health problems (11th ed.)*. https://icd.who.int/en

Wynants, L., van Smeden, M., McLernon, D. J., Timmerman, D., Steyerberg, E. W., & Van Calster, B. (2019). Three myths about risk thresholds for prediction models. *BMC Medicine*, *17*, 192. https://doi.org/10.1186/s12916-019-1425-3

Yamaguchi, S., Foo, J. C., Nishida, A., Ogawa, S., Togo, F., & Sasaki, T. (2020). Mental health literacy programs for school teachers: A systematic review and narrative synthesis. *Early Intervention in Psychiatry*, *14*(1), 14–25. https://doi.org/10.1111/eip.12793

Yamasaki, S., Ando, S., Shimodera, S., Endo, K., Okazaki, Y., Asukai, N., Usami, S., Nishida, A., & Sasaki, T. (2016). The Recognition of Mental Illness, Schizophrenia Identification, and Help-Seeking from Friends in Late Adolescence. *PLOS ONE*, *11*(3), e0151298. https://doi.org/10.1371/journal.pone.0151298

Yan, W.-J., Ruan, Q.-N., & Jiang, K. (2022). Challenges for Artificial Intelligence in Recognizing Mental Disorders. *Diagnostics*, *13*(1), 2. https://doi.org/10.3390/diagnostics13010002

Young, K. S. (1998). Internet addiction: The emergence of a new clinical disorder. *CyberPsychology & Behavior*, *1*(3), 237–244. https://doi.org/10.1089/cpb.1998.1.237

Zimmerman, M., Posternak, M. A., Chelminski, I., & Solomon, D. A. (2004). Using Questionnaires to Screen for Psychiatric Disorders: A Comment on a Study of Screening for Bipolar Disorder in the Community: (Commentary). *The Journal of Clinical Psychiatry*, *65*(5), 605–610. https://doi.org/10.4088/JCP.v65n0503

Annex 1: Per-diagnosis performance for models and sum-scores using only non-proprietary and only parent-report assessments

Only non-proprietary assessments

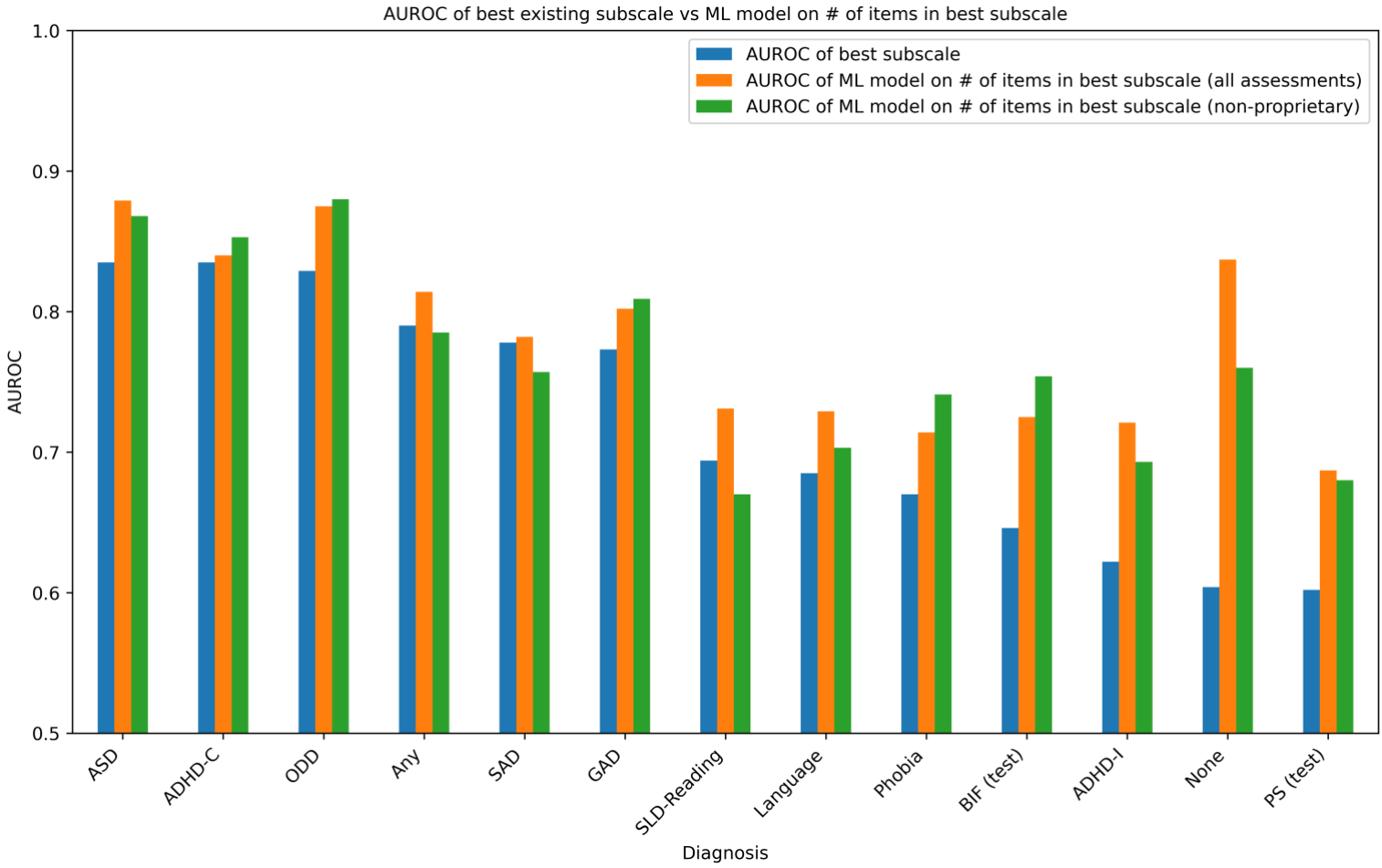

Figure 30. Comparison of the predictive performance of machine learning models using all assessments and only non-proprietary assessments, and the performance of the best subscale (non-learning diagnoses).

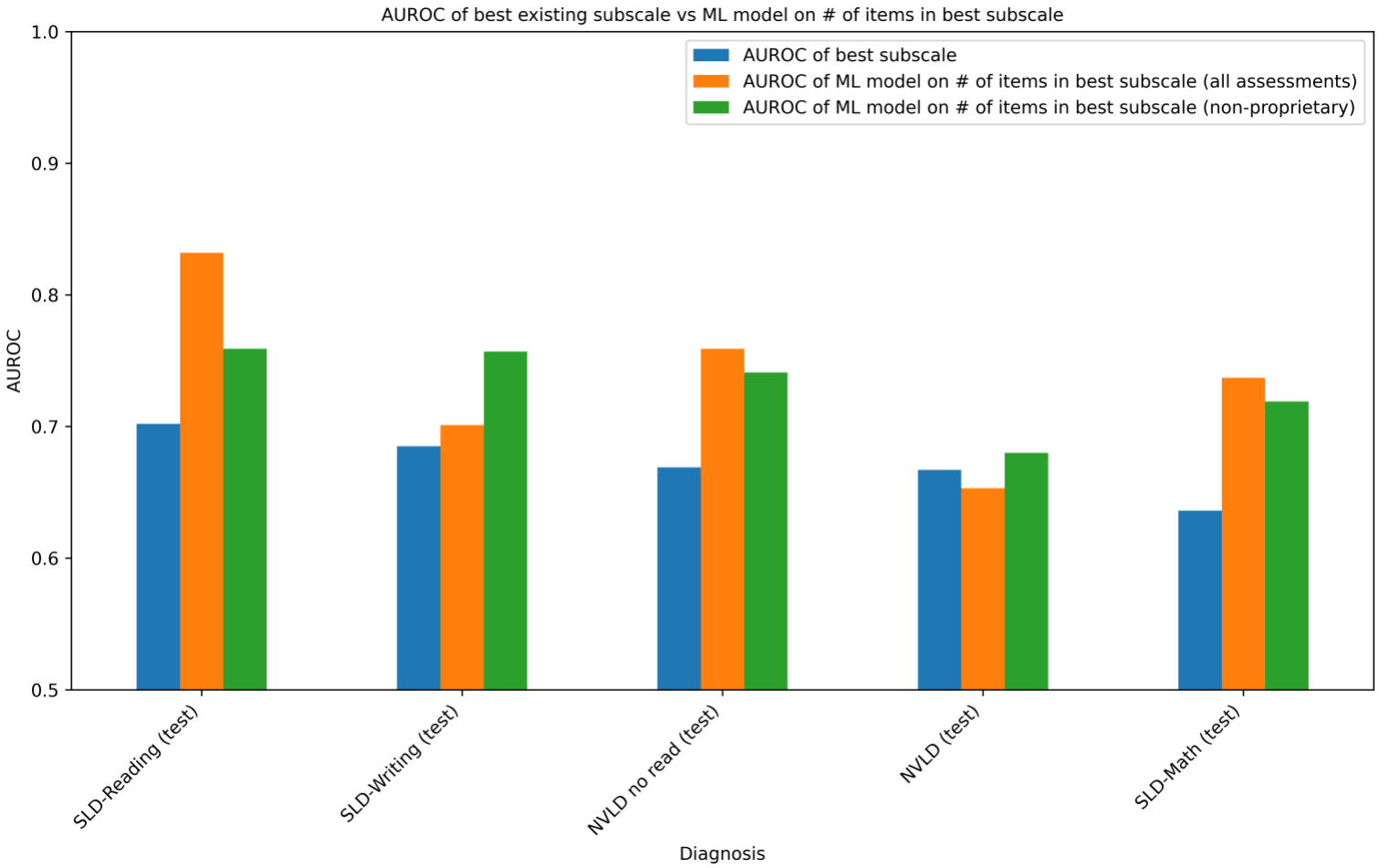

Figure 31. Comparison of the predictive performance of machine learning models using all assessments and only non-proprietary assessments, and the performance of the best subscale (learning diagnoses).

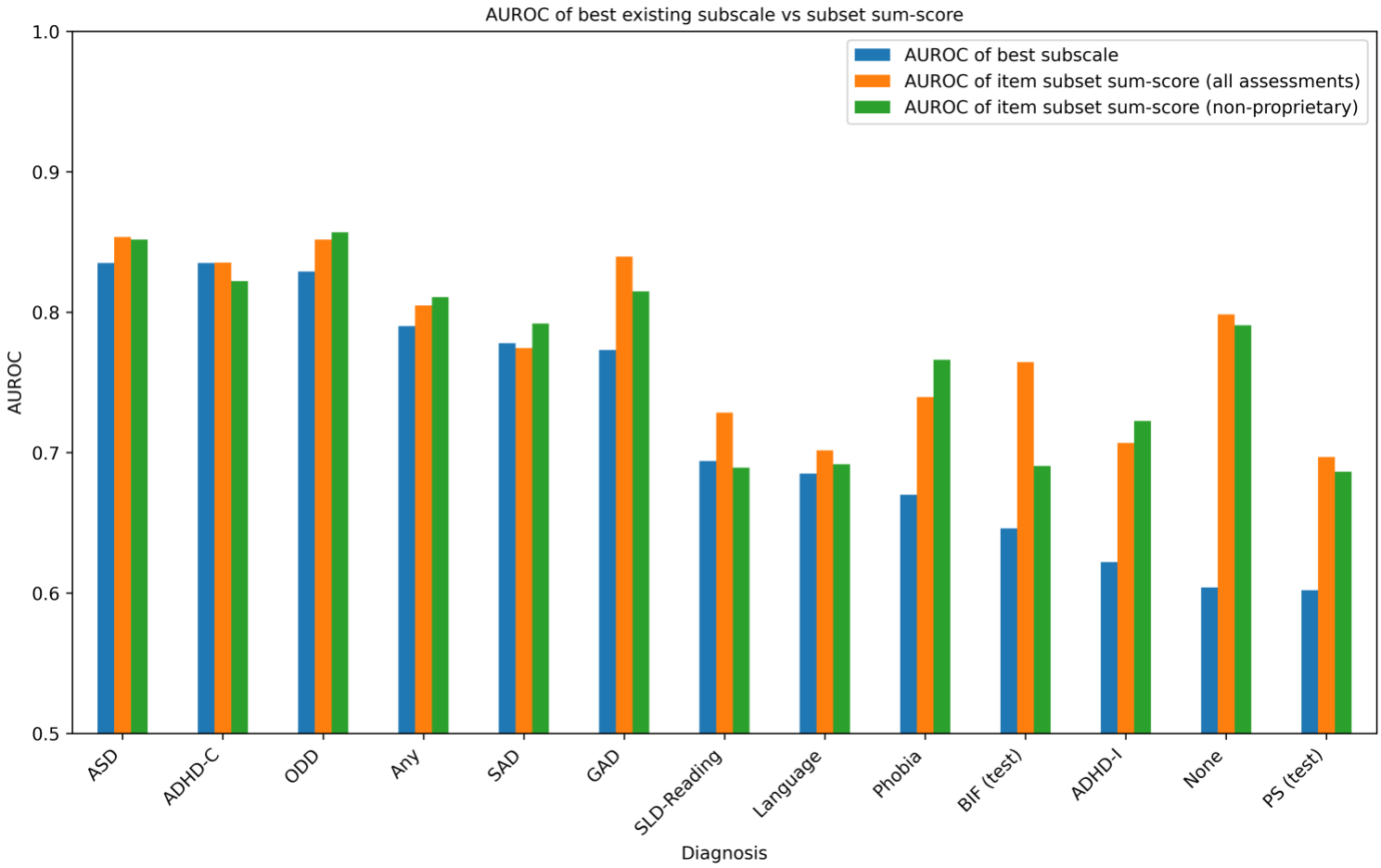

Figure 32. Comparison of the predictive performance of the sum-scores of identified item subsets using all assessments and only non-proprietary assessments, and the performance of the best subscale (non-learning diagnoses).

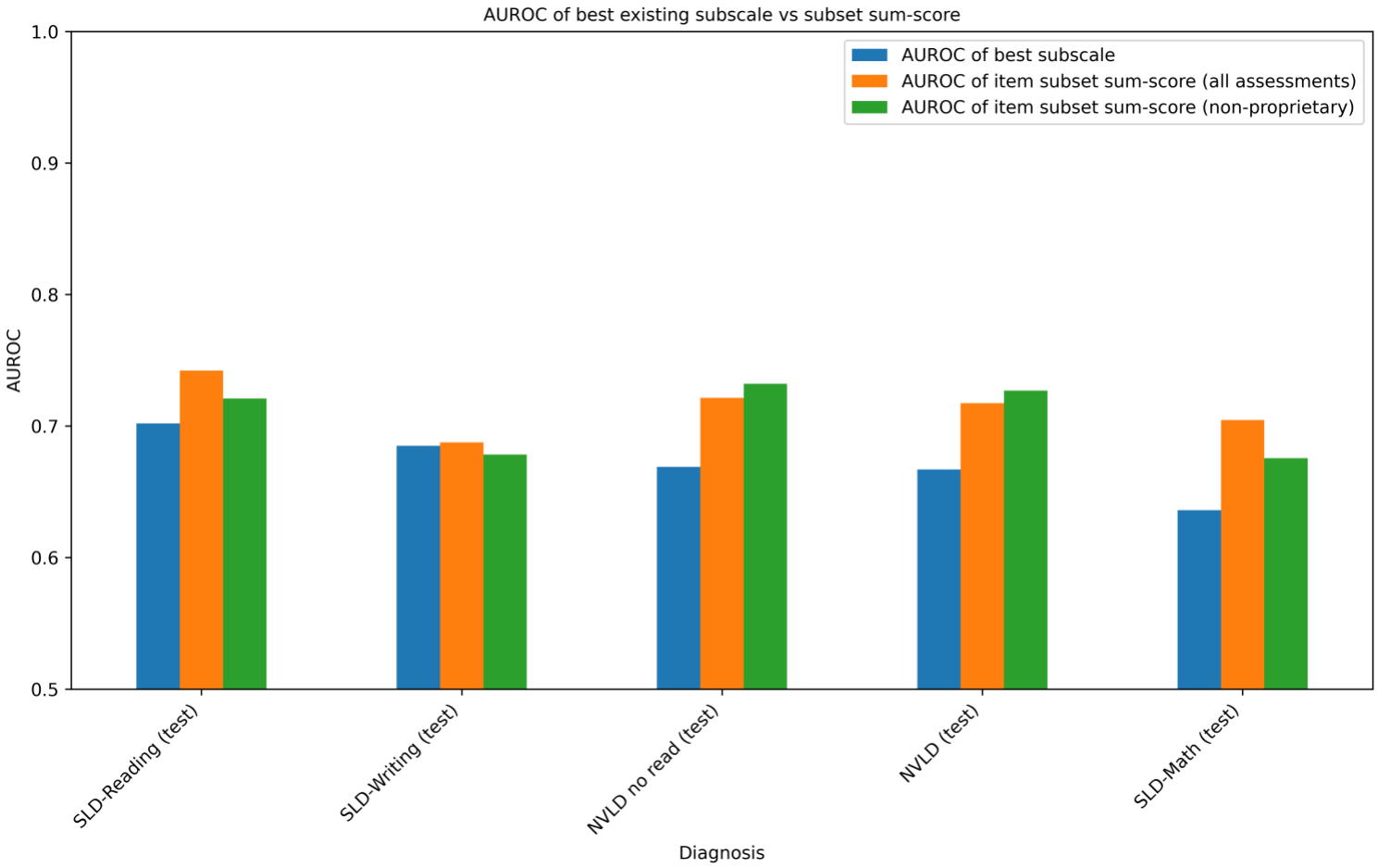

Figure 33. Comparison of the predictive performance of the sum-scores of identified item subsets using all assessments and only non-proprietary assessments, and the performance of the best subscale (learning diagnoses).

Only parent-report assessments

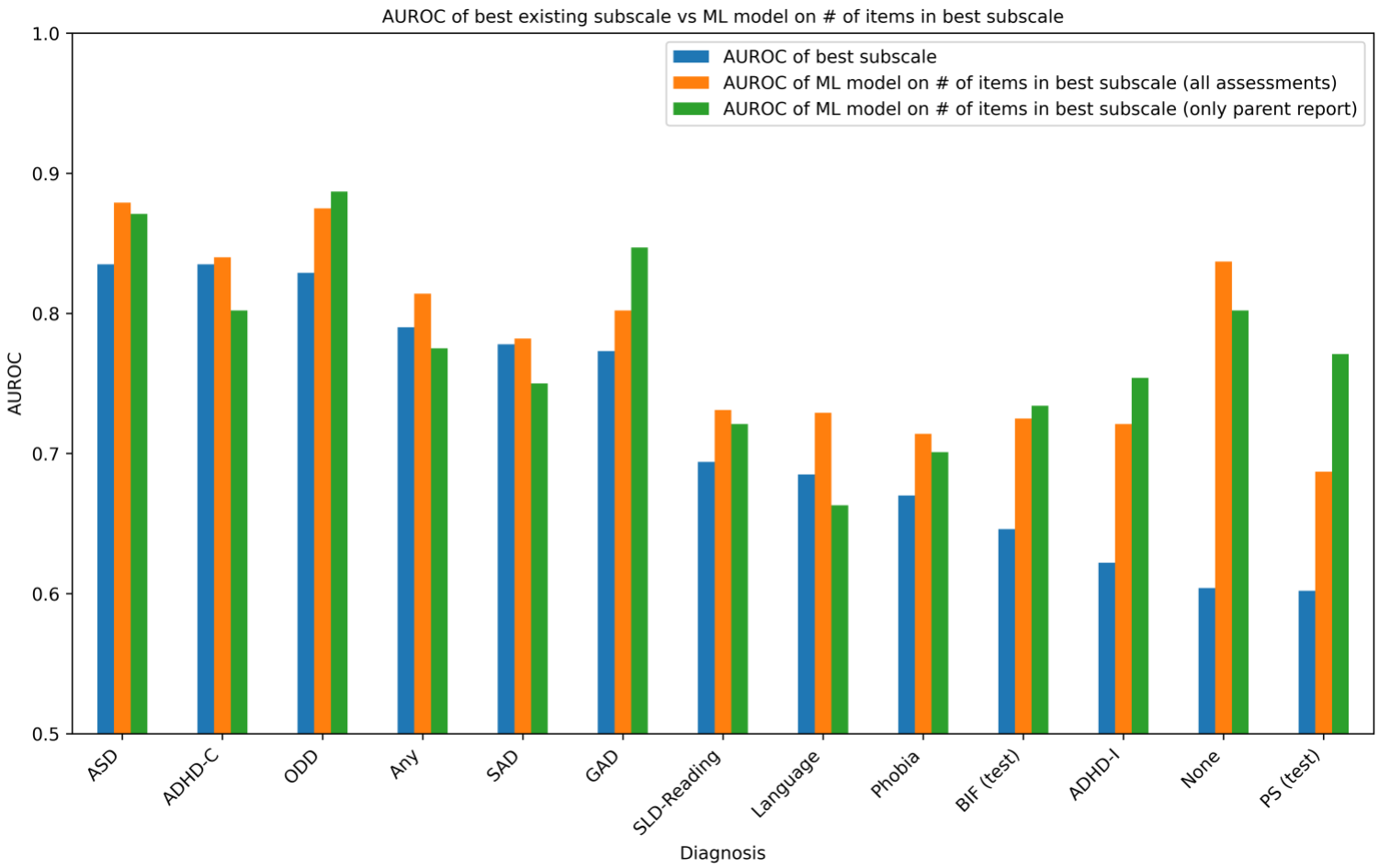

Figure 34. Comparison of the predictive performance of machine learning models using all assessments and only parent-report assessments, and the performance of the best subscale (non-learning diagnoses).

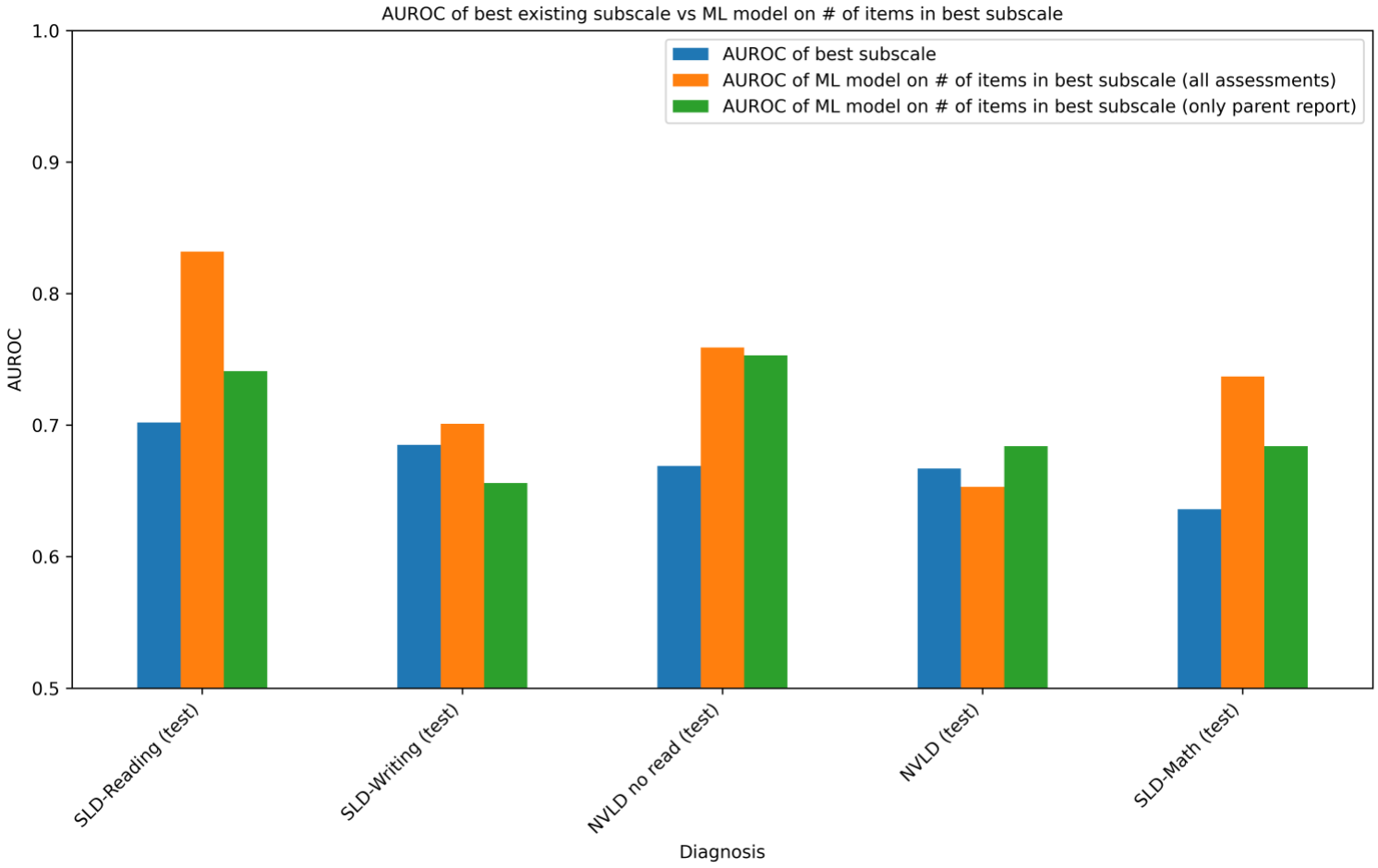

Figure 35. Comparison of the predictive performance of machine learning models using all assessments and only parent-report assessments, and the performance of the best subscale (learning diagnoses).

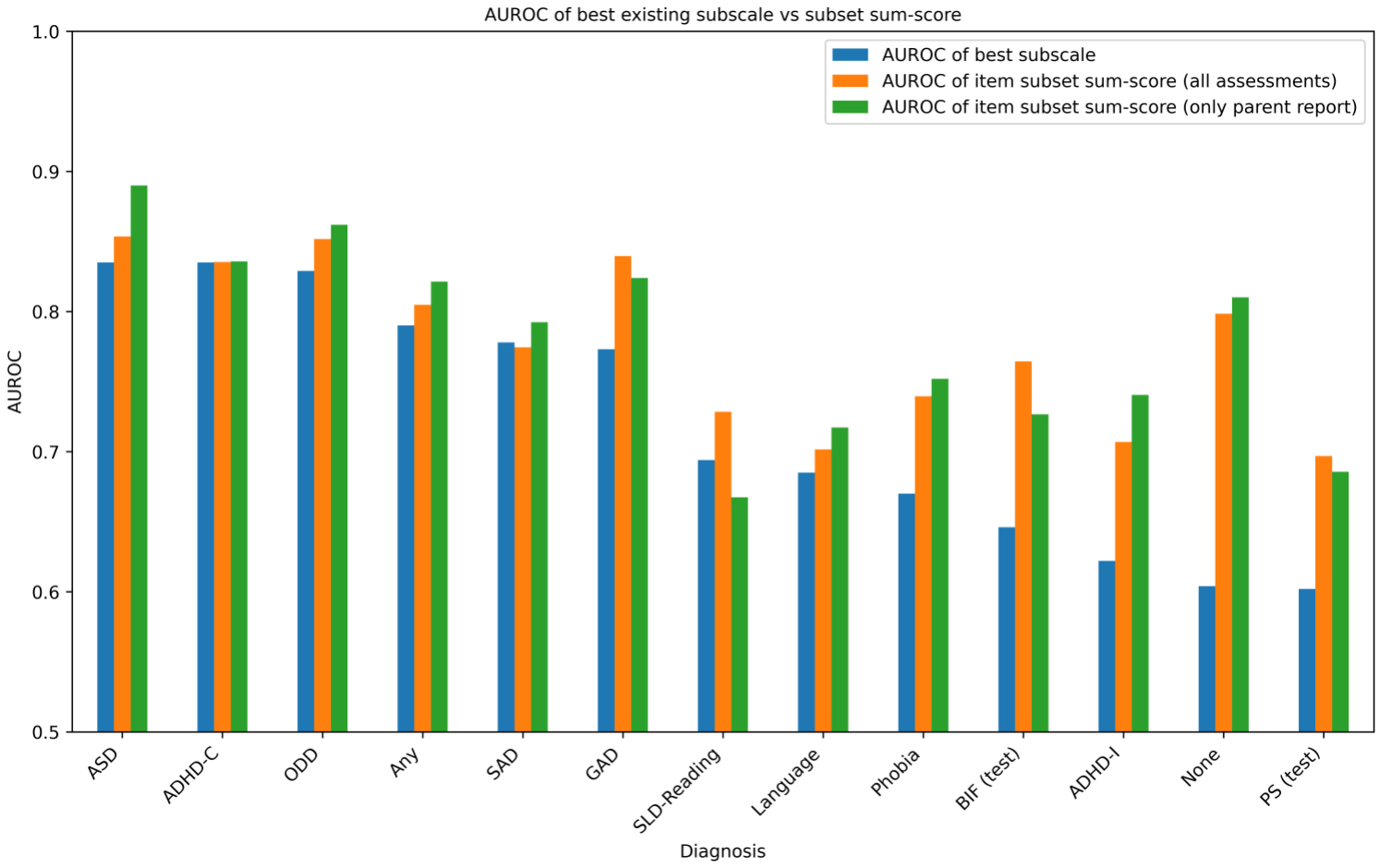

Figure 36. Comparison of the predictive performance of the sum-scores of identified item subsets using all assessments and only parent-report assessments, and the performance of the best subscale (non-learning diagnoses).

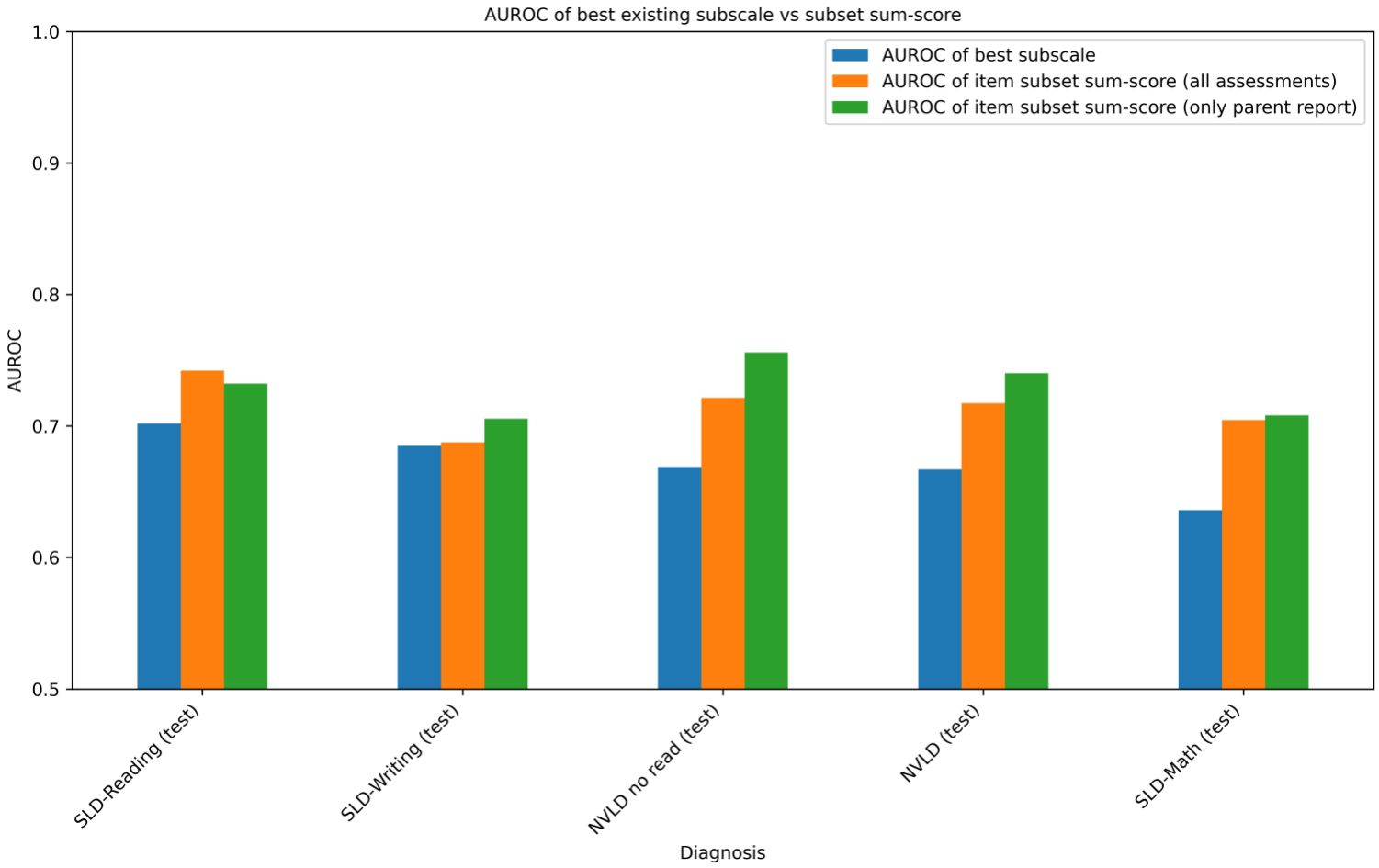

Figure 37. Comparison of the predictive performance of the sum-scores of identified item subsets using all assessments and only parent-report assessments, and the performance of the best subscale (learning diagnoses).

Annex 2: Saturation curves

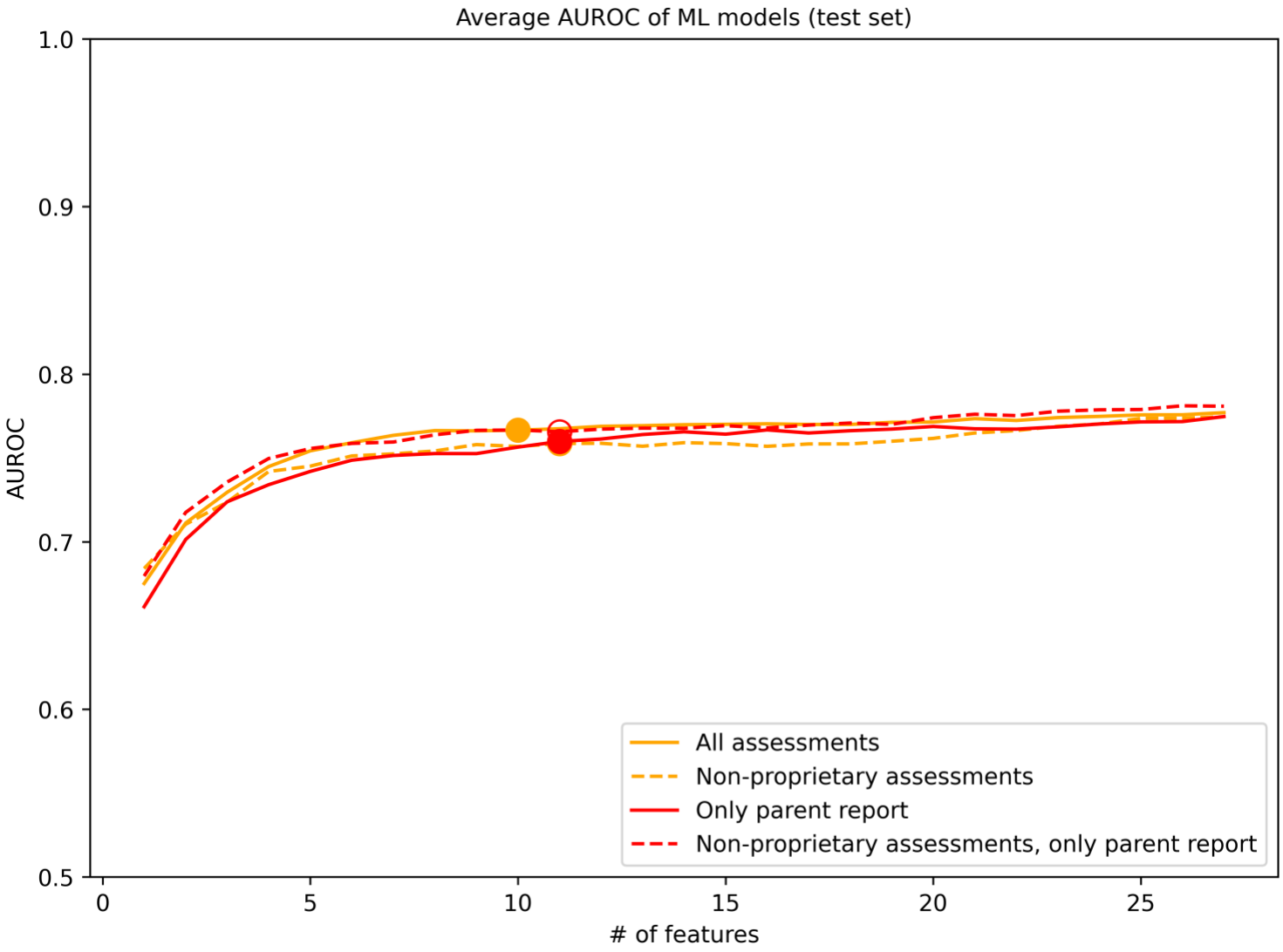

Figure 38. Average performance saturation for the models using different input assessments subsets. The highlighted marker represents the recommended number of features.

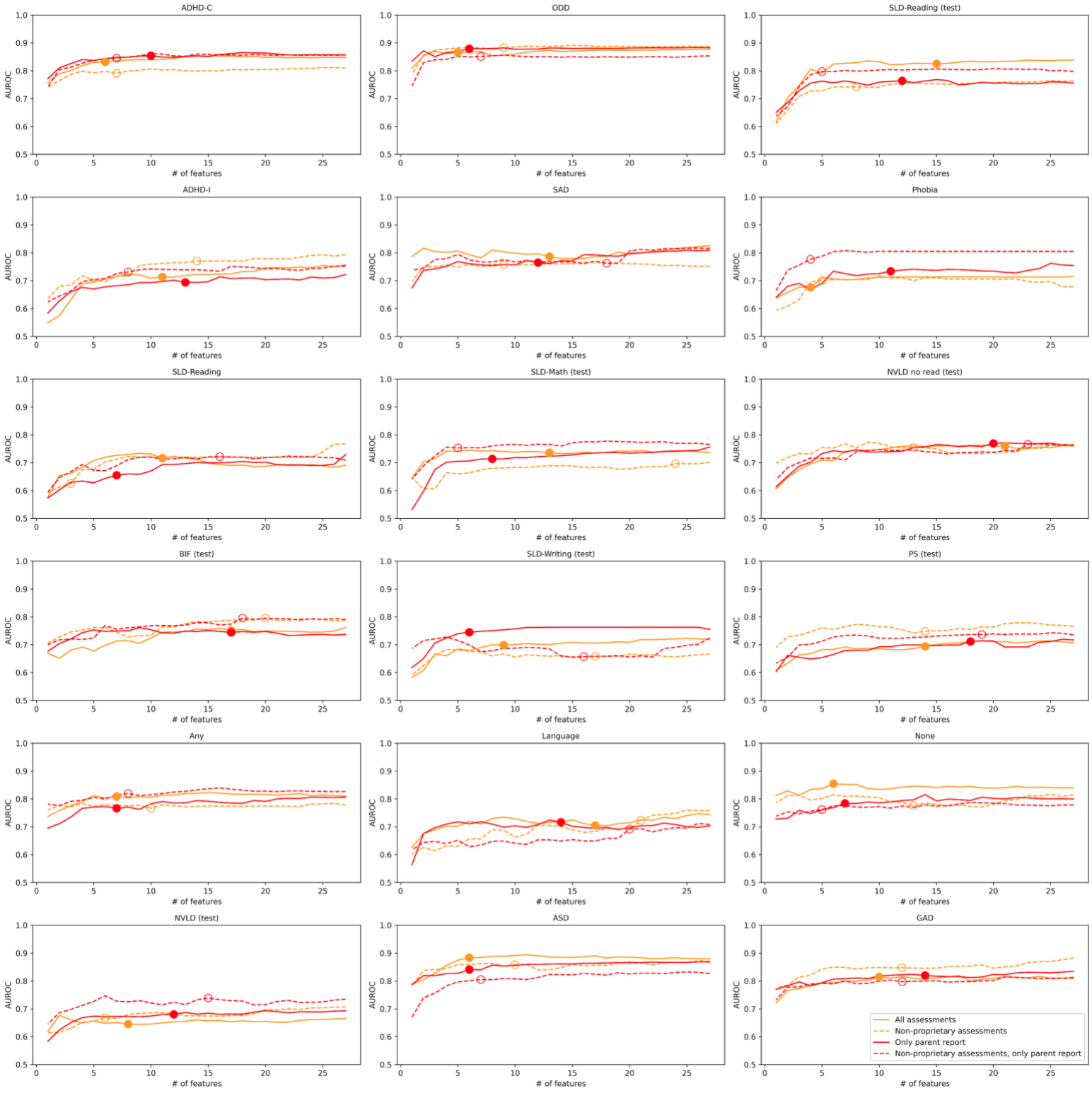

Figure 39. Performance saturation for the models using different input assessments subsets for each diagnosis. The highlighted marker represents the recommended number of features.

Annex 3: Recommended item subsets

Value in parentheses indicates the Elastic Net coefficient.

Learning disorder diagnoses

SLD-Reading (test)

- (0.87) C3SR,C3SR_33: 33. I have trouble with reading. - 0=Not true at all (Never, Seldom), 1=Just a little true (Occasionally), 2=Pretty much true (Often, Quite a bit), 3=Very much true (Very often, Very frequently)
- (0.45) PSI,PSI_18: 18. My child doesn't seem to learn as quickly as most children. - 1=Strongly Disagree, 2= Disagree, 3= Not Sure, 4= Agree, 5= Strongly Agree
- (-0.39) C3SR,C3SR_38: 38. I have trouble with math. - 0=Not true at all (Never, Seldom), 1=Just a little true (Occasionally), 2=Pretty much true (Often, Quite a bit), 3=Very much true (Very often, Very frequently)
- (0.42) PreInt_EduHx,repeated_grades: Were any grades repeated? - 0= No, 1= Yes
- (0.21) APQ_SR,APQ_SR_05: 5. Your parents reward or give something extra to you for behaving well - 1=Never, 2=Almost Never, 3=Sometimes, 4=Often, 5=Always
- (-0.34) SCARED_SR,SCARED_SR_10: 10. I feel nervous with people I don't know well - 0=Not True or Hardly Ever True, 1=Somewhat True or Sometimes True, 2=Very True or Often True

SLD-Writing (test)

- (0.22) C3SR,C3SR_33: 33. I have trouble with reading. - 0=Not true at all (Never, Seldom), 1=Just a little true (Occasionally), 2=Pretty much true (Often, Quite a bit), 3=Very much true (Very often, Very frequently)
- (0.14) PreInt_EduHx,repeated_grades: Were any grades repeated? - 0= No, 1= Yes
- (-0.07) SCARED_SR,SCARED_SR_39: 39. I feel nervous when I am with other children or adults and I have to do something while they watch me (for example: read aloud, speak, play a game, play a sport) - 0=Not True or Hardly Ever True, 1=Somewhat True or Sometimes True, 2=Very True or Often True

NVLD-no-reading (test)

- (0.42) PSI,PSI_18: 18. My child doesn't seem to learn as quickly as most children. - 1=Strongly Disagree, 2= Disagree, 3= Not Sure, 4= Agree, 5= Strongly Agree
- (0.29) PreInt_EduHx,weakness_math: Math - 0= Unchecked, 1= Checked
- (-0.26) Barratt,financialsupport
- (0.19) C3SR,C3SR_09: 9. I have trouble understanding what I read. - 0=Not true at all (Never, Seldom), 1=Just a little true (Occasionally), 2=Pretty much true (Often, Quite a bit), 3=Very much true (Very often, Very frequently)
- (0.23) SRS,SRS_35: 35. Has trouble keeping up with the flow of a normal conversation. - 0= Not True, 1= Sometimes True, 2= Often True, 3= Almost Always True
- (0.18) SCQ,SCQ_09: 9. Does her/his facial expression usually seem appropriate to the particular situation, as far as you can tell? - 1=No, 0=Yes

NVLD (test)

- (0.21) C3SR,C3SR_38: 38. I have trouble with math. - 0=Not true at all (Never, Seldom), 1=Just a little true (Occasionally), 2=Pretty much true (Often, Quite a bit), 3=Very much true (Very often, Very frequently)
- (0.18) RBS,RBS_12: Rubs or scratches marks on arms, leg, face or torso - 0=Behavior does not occur,1=Behavior occurs and is a mild problem,2=Behavior occurs and is a moderate problem,3=Behavior occurs and is a severe problem
- (0.19) PSI,PSI_18: 18. My child doesn't seem to learn as quickly as most children. - 1=Strongly Disagree, 2= Disagree, 3= Not Sure, 4= Agree, 5= Strongly Agree
- (-0.18) PreInt_EduHx,strength_math: Math - 0= Unchecked, 1= Checked
- (0.13) CCSC,CCSC_37: 37. You talked to another adult, other than your parent, who could help you solve the problem. - 1=Never, 2=Sometimes, 3=Often, 4=Most of the time
- (0.13) Basic_Demos,Age
- (0.12) MFQ_SR,MFQ_SR_20: 20. I didn't want to see my friends. - 0=Not True, 1= Sometimes, 2=True
- (0.14) SympChck,CSC_50C: 50. Is preoccupied with very specific objects, routines, or interests(Current) - 0=No, 1=Yes

SLD-Math (test)

- (0.30) PSI,PSI_18: 18. My child doesn't seem to learn as quickly as most children. - 1=Strongly Disagree, 2= Disagree, 3= Not Sure, 4= Agree, 5= Strongly Agree
- (0.25) C3SR,C3SR_38: 38. I have trouble with math. - 0=Not true at all (Never, Seldom), 1=Just a little true (Occasionally), 2=Pretty much true (Often, Quite a bit), 3=Very much true (Very often, Very frequently)
- (-0.30) SCARED_SR,SCARED_SR_40: 40. I feel nervous when I am going to parties, dances, or any place where there will be people that I don't know well - 0=Not True or Hardly Ever True, 1=Somewhat True or Sometimes True, 2=Very True or Often True
- (0.22) Basic_Demos,Age
- (0.21) PreInt_EduHx,recent_grades: Recent typical academic performance: - 1= Excellent, 2= Good, 3= Fair, 4= Poor, 5= Failing
- (0.17) C3SR,C3SR_09: 9. I have trouble understanding what I read. - 0=Not true at all (Never, Seldom), 1=Just a little true (Occasionally), 2=Pretty much true (Often, Quite a bit), 3=Very much true (Very often, Very frequently)
- (0.12) CCSC,CCSC_27: 27. You talked with friends about what you would like to happen. - 1=Never, 2=Sometimes, 3=Often, 4=Most of the time
- (0.15) RBS,RBS_36: Likes the same CD, tape, record or piece of music played continually; Likes same movie / video or part of movie / video - 0=Behavior does not occur,1=Behavior occurs and is a mild problem,2=Behavior occurs and is a moderate problem,3=Behavior occurs and is a severe problem

Non-learning disorder diagnoses

ASD

- (0.08) ASSQ,ASSQ_11: uses language freely but fails to make adjustments to fit social contexts or the needs of different listeners - 0=No, 1=Somewhat, 2=Yes
- (0.07) SRS,SRS_29: 29. Is regarded by other children as odd or weird. - 0= Not True, 1= Sometimes True, 2= Often True, 3= Almost Always True
- (0.07) SRS,SRS_28: 28. Thinks or talks about the same thing over and over. - 0= Not True, 1= Sometimes True, 2= Often True, 3= Almost Always True
- (0.06) SCQ,SCQ_13: 13. Does she/he ever have any special interests that are unusual in their intensity but otherwise appropriate for her/his age and peer group (e.g., trains or dinosaurs)? - 0= No, 1= Yes

- (0.06) SRS,SRS_22: 22. Plays appropriately with children his or her age. - 3= Not True, 2= Sometimes True, 1= Often True, 0= Almost Always True
- (0.05) PreInt_DevHx,temp_11: Problems with social relatedness - 0= No, 1= Yes

ODD

- (0.44) ESWAN,DMDD_8A: 8a. Avoid or limit temper tantrums at home - -3= Far above average,-2= Above average,-1= Slightly above average,0= Average,1= Slightly below average,2= Below average,3= Far below average
- (0.33) SympChck,CSC_39P: 39. Argues or talks back to adults, more than others his/her age(Past) - 0=No, 1=Yes
- (0.33) CBCL,CBCL_22: 22. Disobedient at home - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (0.18) SympChck,CSC_40P: 40. Actively disobeys or doesn't listen to adult rules(Past) - 0=No, 1=Yes
- (0.21) SympChck,CSC_05C: 5. Has strong and explosive feelings of anger(Current) - 0=No, 1=Yes

GAD

- (0.47) SDQ,SDQ_08: Many worries or often seems worried - 0=Not True, 1=Somewhat True, 2=Certainly True
- (-0.46) ESWAN,Panic_A01B_WAS_MISSING
- (0.23) SCARED_P,SCARED_P_07: 7. My child is nervous - 0=Not True or Hardly Ever True, 1=Somewhat True or Sometimes True, 2=Very True or Often True
- (-0.24) CBCL,CBCL_111: 111. Withdrawn, doesn't get inolved with others - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (-0.15) APQ_P,APQ_P_03: 3. You threaten to punish your child and then do not actually punish him/her - 1=Never, 2=Almost Never, 3=Sometimes, 4=Often, 5=Always
- (0.23) ESWAN,Panic_A02A
- (0.17) PCIAT,PCIAT_07: 7. How often does your child check his or her e-mail before doing something else? - 0=Does Not Apply, 1=Rarely, 2=Occasionally, 3=Frequently, 4=Often, 5=Always
- (0.19) SCARED_P,SCARED_P_23: 23. My child is a worrier - 0=Not True or Hardly Ever True, 1=Somewhat True or Sometimes True, 2=Very True or Often True
- (0.13) Basic_Demos,Sex

- (0.19) SCARED_P,SCARED_P_35: 35. My child worries about how well he/she does things - 0=Not True or Hardly Ever True, 1=Somewhat True or Sometimes True, 2=Very True or Often True

ADHD-C

- (0.45) CBCL,CBCL_10: 10. Can't sit still, restless or hyperactive - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (0.55) SWAN,SWAN_16: 16. Reflects on questions (controls blurting out answers) - -3= Far above average,-2= Above average,-1= Slightly above average,0= Average,1= Slightly below average,2= Below average,3= Far below average
- (0.37) SympChck,CSC_36P: 36. Has difficulty remaining seated at home or school(Past) - 0=No, 1=Yes
- (-0.41) ESWAN,Panic_A03B_WAS_MISSING
- (0.25) SWAN,SWAN_11: 11. Stays seated (when required by class rules or social conventions) - -3= Far above average,-2= Above average,-1= Slightly above average,0= Average,1= Slightly below average,2= Below average,3= Far below average
- (0.19) CBCL,CBCL_74: 74. Showing off or clowning - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true

Language

- (0.19) SCQ,SCQ_05: 5. Does she/he ever get her/his pronouns mixed up (e.g., saying you or she/he for I)? - 0= No, 1= Yes
- (0.15) PreInt_EduHx,repeated_grades: Were any grades repeated? - 0= No, 1= Yes
- (0.14) SympChck,CSC_21P: 21. Is unable to speak in specific situations, such as school, despite being able to speak without a problem in other situations(Past) - 0=No, 1=Yes
- (0.11) ARI_S,ARI_S_04: I am angry most of the time - 0=Not True, 1=Somewhat True, 2=Certainly True
- (-0.08) APQ_P,APQ_P_34: 34. You ignore your child when he/she is misbehvaing - 1=Never, 2=Almost Never, 3=Sometimes, 4=Often, 5=Always
- (0.08) APQ_SR,APQ_SR_28: 28. You stay out later than you are supposed to and your parents don't know it - 1=Never, 2=Almost Never, 3=Sometimes, 4=Often, 5=Always

- (-0.07) MFQ_P,MFQ_P_08: 8. S/he felt s/he was no good anymore. - 0= Not True, 1= Sometimes, 2= True
- (-0.02) DTS,DTS_12: 12. My feelings of distress or being upset scare me. - 1=Strongly Disagree, 2=Mildly Disagree, 3=Agree and Disagree Equally, 4=Mildly Agree, 5=Strongly Agree
- (0.07) ASSQ,ASSQ_26: has markedly unusual facial expression - 0=No, 1=Somewhat, 2=Yes
- (-0.07) CBCL,CBCL_09: 9. Can't get his/her mind off certain thoughts; obsessions - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (0.09) CBCL,CBCL_61: 61. Poor school work - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (0.06) APQ_P,APQ_P_13: 13. You compliment your child when he/she has done something well - 1=Never, 2=Almost Never, 3=Sometimes, 4=Often, 5=Always
- (-0.07) PreInt_DevHx,m_birthage: Mother's age at birth of child
- (0.05) ARI_S,ARI_S_02: I often lose my temper - 0=Not True, 1=Somewhat True, 2=Certainly True
- (-0.06) APQ_P,APQ_P_31: 31. The punishment you give your child depends on your mood - 1=Never, 2=Almost Never, 3=Sometimes, 4=Often, 5=Always
- (-0.06) PreInt_EduHx,strength_english: English - 0= Unchecked, 1= Checked
- (0.05) APQ_SR,APQ_SR_01A_WAS_MISSING

Phobia

- (0.15) SympChck,CSC_22P: 22. Has intense fears of specific animals, situations, or anything else(Past) - 0=No, 1=Yes
- (0.12) SDQ,SDQ_24: Many fears, easily scared - 0=Not True, 1=Somewhat True, 2=Certainly True
- (0.13) CBCL,CBCL_29: 29. Fears certain animals, situations, or places, other than school - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (0.10) CBCL,CBCL_112: 112. Worries - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true

ADHD-I

- (0.05) SWAN,SWAN_05: 5. Organizes tasks and activities - -3= Far above average,-2= Above average,-1= Slightly above average,0= Average,1= Slightly below average,2= Below average,3= Far below average
- (0.04) SympChck,CSC_35P: 35. Is often easily distracted(Past) - 0=No, 1=Yes

- (-0.03) ESWAN,Panic_A02A_WAS_MISSING
- (0.04) CBCL,CBCL_17: 17. Daydreams or gets lost in his/her thoughts - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (-0.03) CBCL,CBCL_19: 19. Demands a lot of attention - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (-0.03) ESWAN,Panic_A02B_WAS_MISSING
- (0.04) SDQ,SDQ_25: Good attention span, sees chores or homework through to the end - 2=Not True, 1=Somewhat True, 0=Certainly True
- (-0.03) SympChck,CSC_38P: 38. Often becomes really upset and loses his/her temper(Past) - 0=No, 1=Yes
- (-0.02) CBCL,CBCL_93: 93. Talks too much - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (-0.02) CBCL,CBCL_28: 28. Breaks rules at home, school, or elsewhere - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (0.03) PCIAT,PCIAT_06: 6. How often do your child's grades suffer because of the amount of time he or she spends online? - 0=Does Not Apply, 1=Rarely, 2=Occasionally, 3=Frequently, 4=Often, 5=Always

SAD

- (0.15) SCARED_P,SCARED_P_41: 41. My child is shy - 0=Not True or Hardly Ever True, 1=Somewhat True or Sometimes True, 2=Very True or Often True
- (0.13) SCARED_P,SCARED_P_40: 40. My child feels nervous when he/she is going to parties, dances, or any place where there will be people that he/she doesn\u00e2\u0080\u0099t know well - 0=Not True or Hardly Ever True, 1=Somewhat True or Sometimes True, 2=Very True or Often True
- (-0.09) SWAN,SWAN_16: 16. Reflects on questions (controls blurting out answers) - -3= Far above average,-2= Above average,-1= Slightly above average,0= Average,1= Slightly below average,2= Below average,3= Far below average
- (0.10) CBCL,CBCL_71: 71. Self-conscious or easily embarrassed - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (-0.07) CBCL,CBCL_07: 7. Bragging, boasting - 0=Not true, 1=Somewhat or sometimes true, 2=Very true or often true
- (-0.06) SympChck,CSC_37C: 37. Often acts before thinking(Current) - 0=No, 1=Yes
- (-0.06) SympChck,CSC_41P: 41. Frequently lies(Past) - 0=No, 1=Yes

- (-0.06) SympChck,CSC_04C: 4. Often feels overly happy and silly, above and beyond a normal feeling(Current) - 0=No, 1=Yes

Annex 4. Original version of FACETS (then "SRQ")

This is a broad range screening questionnaire initially designed for and with school teachers. The purpose of the SRQ is to screen for mental health and learning problems so that teachers can make timely, early or preventive interventions as well as referrals for specialty services.

This questionnaire is intended to be completed by teachers seeking to better understand a students' profile (strengths and needs) and support them accordingly. It is constructed with the help of field experts (clinicians and researchers), combining items from validated questionnaires.

There are no right or wrong answers –.even if answers to some questions are unknown or not applicable, please answer to the best of your ability and add a comment at the end of the questionnaire referring to that particular question. No one behavior is good or bad on its own – we look for combinations/patterns among different behaviors to have more comprehensive student profiles.

Most questions are designed in a manner where both ends (0=left and 10=right response options) are problematic, and the middle is usually the average. For example, paying too little attention and paying too much attention to the extent where the student refuses to do anything else are both problematic.

If a student exhibits the same behavior differently in different settings (e.g., follows rules in class but does not follow rules in the playground), please select one response that you feel is more appropriate, and add the different scenarios as comments at the end. Please do not not mention any identifiable features of the students (e.g. name) in the comment field.

Once the questionnaire is answered, the intent is to provide the teachers a visual interactive dashboard of student learner profiles and recommend classroom strategies and interventions that the teacher can implement to best support the student.

This assessment is not a diagnostic tool. If you suspect an underlying condition or disorder, please refer to a trained mental health professional.

Try to answer all the questions to the best of your knowledge.

<teacher/parent ID recorded upon login>
Student birth year:
Student sex:

COMMUNICATION	
Language (incl Gestures)	(No communication…Imprecise)
Speech Quantity	(Never speaks…Excessive speaking)
Speech Quality	(Never understandable…Imprecise)
SOCIAL FUNCTION	
Social Interaction	(None…Imprecise)
Social Reciprocity	(None…Imprecise)
Social Communication	(None…Imprecise)
Cooperation	(Uncooperative…Overly cooperative)
Bullying	(Ignores…Actively involved)
BEHAVIOR	
Self-Control	(Overly inhibited…Uninhibited)
Compliance	(Never complies…Rigid adherence)
Persistent thoughts	(None…Disruptive)
Habits	(No habits/routines…Disruptive)
Anger management	(Never angered…Intense outbursts)
Aggression	(Submissive…Overly aggressive)
Risk Taking	(Overly cautious…Reckless)
Activity Level	(Inactive…Hyperactive)
Harm to Self/Others	(Denies thought…Harms self/others)
Substance Use	(Denies risk…Disruptive)
Play	(None…Disruptive)
EMOTION	
Mood	(Apathetic…Disruptive)
Worries	(Always carefree…Worries all the time)
Trauma	(Denies…Disruptive)
Attachment	(Lacking attachments…Overly attached)

Empathy	(Insensitive…Overly sensitive)
Irritability	(Placid…Persistently irate)
Emotional Reactivity	(Apathetic…Overly emotional)
Stress Management	(Non-reactive…Overly reactive)
PERSONALITY	
Temperament	(Overly easy-going …Difficult)
Confidence	(None…Excessive)
Creativity	(None…Disruptive)
Responsibility	(Irresponsible…Overly responsible)
Integrity	(Dishonest…Overly honest)
Perseverance	(Unmotivated…Overly tenacious)
Future Outlook	(Pessimistic…Overly optimistic)
COGNITION	
Attention/Concentration	(Unfocused…Overly focused)
Planning and Organization	(Disorganized…Overly organized)
Memory	(Never remembers…Memorizes all detail)
Abstract Thinking	(Overly concrete…Overly abstract)
Thinking speed	(Slow tempo…Disruptively fast)
Self-Image	(Negative…Overly positive)
Perception of Reality	(Denies…Intrusive/disruptive)
Task Completion	(None…Perfectionistic)
LEARNING	
Reading	(No reading…Perfectionistic)
Writing	(No writing…Perfectionistic)
Math	(No math…Perfectionistic)
Learning Strategies	(Does not learn…Perfectionistic)
Academic Performance	(Under-achiever…Perfectionistic)
HEALTH	
Eating	(Restrictive/rigid…Excessive)
Sleep	(Limited/minimal…Excessive)
Menses	(Ignores/denies…Disruptive)
Self-Care/Hygiene	(None…Disruptive)
Sexual Behavior	(Ignores/denies…Disruptive)

Complains of Pain	(Ignores/denies…Disruptive)
SOMATIC FUNCTION	
Gross Motor Skills	(None…Uncoordinated)
Fine Motor Skills	(None…Uncoordinated)
Atypical Movements	(Ignores/denies…Disruptive)
Bodily Responses	(None…Disruptive)
DAILY ROUTINES	
Toileting	(None…Excessive)
Morning Routines	(None…Excessive/disruptive)
After-School Routines	(None…Excessive/disruptive)
Going Out Routines	(None…Excessive/disruptive)
Screen Time	(None …Excessive/disruptive)

Annex 5. Current version of FACETS (English version)

FACETS

(Functional Activity, Cognition, Emotion and Thinking Scale)

FACETS is designed to help identify a child's or adolescent's strengths and needs by assessing multiple dimensions of functioning.

Each FACETS item represents a dimensional view of a particular area of functioning, in an attempt to capture problems or needs at the extremes, and typically expected functioning, or strengths, near the center of the dimensional scale.

There are no right or wrong answers. Responses do not indicate a "good" or "bad" judgment about the person. An individual's profile is a representation of overall functioning.

+----------+----------+

Atypical Strength/Typically Expected Atypical

TYPICALLY EXPECTED

For the purposes of FACETS, "typically expected" represents the distinctive qualities for an individual child's functioning along a particular dimension. "Typically Expected" is neither "average" nor "normal," each of which represents statistical phenomena; instead "typically expected" is the range of developmentally expected functions which are strengths because they represent the abilities most suitable for successful adaptation, in the broadest range of environments and situations. As an example, a "typically expected" rating for gross motor skills represents a child's ability to use motor skills for participation in all tasks and activities appropriate for their age - whether they are a future professional athlete or a casual sports participant.

ATYPICAL

For the purposes of FACETS, "Atypical" represents the lack of distinctive qualities for an individual's functioning along a particular dimension. As an example, an "atypical" rating for gross motor skills, at either extreme, represents a child's inability to use motor skills for participation in any tasks and activities appropriate for their age.

Instructions

When completing FACETS, please:

1. Choose the “best estimate” for each item as is appropriate for this child.
2. Rate the child in a manner that reflects the overall functioning, in all settings; if the child functions differently in different settings, provide an aggregate rating or general impression.
3. Rate all items to the best of your ability, even for items that may not seem to be applicable.
4. Add comments at the end of the questionnaire, if you wish.

FACETS is not a diagnostic tool. If you suspect an underlying clinical condition or disorder, please consult with a qualified health professional.

FACETS

(Functional Activity, Cognition, Emotion and Thinking Scale)

1) COMMUNICATION

A. Expressive Language — communicating with speech and gestures

1. Does not use words or gestures
2. Uses words and gestures to effectively communicate
3. Uses words and gestures but communicates ineffectively

B. Receptive Language – recognizing and understanding words and gestures as communication

1. Does not recognize words and gestures
2. Understands words and gestures
3. Recognizes words and gestures but cannot interpret meaning

C. Speech Quantity – speaking with an appropriate number of words for communication

1. Never uses words to communicate
2. Uses appropriate number of words to communicate
3. Speaks excessively

D. Speech Quality – speaking understandably

1. Speech not understandable
2. Speech clear and articulate
3. Speech overly precise and excessively perfectionistic

2) SOCIAL FUNCTION

A. Social Engagement — engaging in social behavior

1. Does not engage socially
2. Actively and adaptively socially engaged
3. Engages but socially inappropriate

B. Social Communication — communication enabling social interactions
1. Lacks communication necessary for social interactions
2. Communication facilitates reciprocal and adaptive social behavior
3. Communication lacks reciprocity and social appropriateness

C. Social Cooperation — cooperating in social interactions, including play
1. Uncooperative in work, school, play, and/or other reciprocal social interactions
2. Cooperates flexibly and appropriately in work, school, play, and/or other reciprocal social interactions
3. Overly cooperative, even when disadvantageous, in work, school, play, and/or other reciprocal social interactions

D. Tolerance – ability to appropriately respond to different people, cultures, values, and belief systems
1. Denies or fails to recognize differences in cultures and belief systems
2. Appreciates and/or responds appropriately and adaptively to individuals and/or ideas from other cultures or belief systems
3. Hostile and refuses to acknowledge or accept individuals and/or ideas from any different culture or belief system

3) BEHAVIOR
A. Self-control — managing impulses and self-regulating behavior
1. Excessively restricted/inhibited behavior
2. Balances impulses and uses self-regulation adaptively
3. Excessively uninhibited behavior

B. Compliance — following rules and instructions adaptively
1. Never follows rules/instructions
2. Follows rules/instructions adaptively
3. Excessively rigid in following rules/instructions

C. Obsessive Thoughts — managing recurring and/or persistent thoughts, ideas, and/or interests
1. Unable to use recurring and/or persistent thoughts, ideas, and/or interests to achieve goals
2. Manages recurring and/or persistent thoughts, ideas, and/or interests adaptively to achieve goals
3. Recurring and/or persistent thoughts/ideas are obsessive, intrusive and disrupt daily function

D. Habits/routines — managing recurring habits
1. No useful habits/routines present
2. Habits/routines useful and adaptive
3. Compulsive, rigid habits/routines that interfere with daily functioning

E. Anger Management — managing responses when provoked/frustrated
1. No responses when provoked/frustrated
2. Appropriate response when provoked/frustrated
3. Rage, temper tantrums, violent behavior when provoked/frustrated

F. Assertiveness — initiating actions to support interests, goals, and desires
1. Does not take assertive actions
2. Appropriately assertive
3. Overly assertive, pushy, or aggressive

G. Risk Taking — takes chances to achieve a goal when faced with possible failure, embarrassment, or harm
1. Overly cautious; refuses to take any chances
2. Appropriately takes risks to achieve goals
3. Reckless, dangerous, takes excessive risks

H. Activity Level — maintaining physical and cognitive activity
1. Inactive or sedentary
2. Maintains appropriate activity level
3. Hyperactivity disrupting adaptive functioning

I. Plays by Self and with Others — playing alone and with others
1. Does not play alone or with others
2. Appropriately and flexibly engages in and sustains play alone and with others
3. Disruptive/disorganized play alone and with others

4) EMOTION
A. Emotional regulation and reactivity – managing threshold and intensity of emotional responses
1. Apathetic, no apparent moods or feelings
2. Adaptively regulates emotions and emotional responses
3. Extreme or disruptive moods, feelings, or emotional responses

B. Worries/Anxiety — managing worries and anxiety
1. Always carefree, never worries, in any situation
2. Adaptive, functional, well-regulated worries and anxiety
3. Persistently anxious, worries about everything

C. Attachment — having attachments to family, peers, and adults
1. No attachments to others
2. Adaptive, flexible, differentiated attachments to others
3. Inappropriately or overly attached to others

D. Empathy — having sensitivity to the feelings and needs of others
1. Insensitive to feelings or needs of others
2. Adaptively, flexibly sensitive to feelings and needs of others
3. Hypersensitive to feelings or needs of others

E. Irritability — reacting to frustration or annoyance
1. Placid or indifferent; never frustrated/upset

2. Adaptively reacts when frustrated or annoyed
3. Easily and persistently frustrated or upset

F. Stress Management — managing tension resulting from challenges or demands
1. No tension in response to difficulties or demands
2. Adaptively and flexibly manages tension in response to difficulties and demands
3. Overreacts with excessive tension in response to major or minor difficulties and demands

5) PERSONALITY
A. Self-Confidence - assessing one's belief in their ability
1. No self-confidence
2. Appropriately and flexibly assesses own ability
3. Overly confident; so sure of self that leads to mistakes

B. Creativity — applying novel strategies to problem-solving
1. No imagination and creativity
2. Flexibly applies novel strategies in problem-solving
3. Extravagant thoughts or ideas interfering with problem-solving

C. Responsibility for self and others — accepting responsibility for actions taken by self or others
1. Refuses or is unable to take any responsibility
2. Appropriately accepts responsibility for self and others
3. Tries to assume too much responsibility for self and others

D. Integrity and Honesty — recognizing and sharing the difference between truth and false with others
1. Unable or unwilling to assess the difference between true and false
2. Knows the difference between true and false, and shares it appropriately with others
3. Dishonest; lies or misrepresents the truth

E. Perseverance — persisting when undertaking tasks
1. Does not try even the simplest tasks
2. Flexibly persists in attempting appropriate tasks
3. Stubbornly refuse to consider other strategies when facing repeated failure

F. Future Outlook — having thoughts and feelings about the future
1. Persistently pessimistic, unrealistically negative future outlook
2. Appropriately balanced future outlook
3. Persistently optimistic, unrealistically positive future outlook

6) COGNITION
A. Attention/Concentration - regulating attention and concentration
1. Does not pay attention or concentrate
2. Flexibly and adaptively regulates attention and concentration
3. Overly focused; difficulty changing attention interferes with overall functioning

B. Planning and Organization - developing plans to complete tasks
1. Disorganized; does not plan
2. Adaptive, flexible planning
3. Rigid, unrealistic, or over-planning

C. Memory/Recall - expending effort and employing strategies for remembering and recalling
1. Does not remember or recall even essential information
2. Can prioritize and use memory tools (mnemonics) for memory or recall
3. Tries to, or does, memorize everything, without differentiating essential from non-essential, interfering with overall functioning

D. Abstract Thinking - understanding and applying abstract concepts, such as analogies and metaphors
1. Does not understand abstract concepts, including the implied meaning of words and expressions
2. Adaptively and practically interprets abstract concepts, such as metaphors
3. Understands all concepts but can not practically put ideas into actions

E. Thinking Speed - adapting thinking speed to meet situational demands
1. Always thinks slowly and inefficiently
2. Modulates and adapts thinking speed to optimize comprehension and output
3. Excessively fast thinking leading to conceptual and adaptive errors, and misunderstanding

F. Self-Image — mental picture of one's own attributes
1. Highly self-critical, excessively negative
2. Appropriately recognizes one's own attributes
3. Self-aggrandizing, excessively positive

G. Task Completion — recognizing beginning, structure, and end of tasks
1. Never recognizes beginning and end of tasks
2. Completes tasks fully in a timely manner
3. Disruptively perfectionistic, interfering with task completion

7) LEARNING and ACQUISITION of knowledge and skills
A. Reading — cognitive and mechanical elements of reading comprehension
1. Unable to recognize letters, words, and/or grammar for reading comprehension
2. Reading is efficient, effective, and developmentally/age appropriate
3. Reading disrupted by excessive speed and/or meticulousness

B. Writing — cognitive and mechanical elements of written expression
1. Unable to use letters, words, and/or grammar for written expression
2. Writing is efficient, effective, and developmentally/age appropriate
3. Writing disrupted by excessive speed and/or meticulousness

C. Mathematics — cognitively and mechanically performing mathematical operations

1. Unable to understand numbers and variables, and/or to perform mathematical operations and/or processes
2. Use of numbers, variables, and mathematical operations and/or processes is efficient, effective, and age/developmentally appropriate
3. Mathematical operations and processes disrupted by excessive speed and/or meticulousness

D. Learning Strategies — adopting and developing strategies for learning new material
1. Does not have any strategy for learning
2. Develops and adopts adaptive strategies for learning
3. Learning strategies disrupted by rigidity, excessive speed, and/or meticulousness

E. Academic Motivation — willingness to apply skills and abilities to achieve academic goals
1. No interest or desire to try and engage in academic activities
2. Self-starter who adaptively uses skills and abilities to achieve academic goals
3. Excessive preoccupation with academic achievement leading to stress and disrupted performance and learning

F. Gross Motor Skills — strength and coordination of motor skills, e.g., walking, running and jumping
1. Low muscle tone with sluggishness and impaired function
2. Gross motor activity is efficient, effective, and developmentally/age appropriate
3. Excessive muscle tone, muscle stiffness, impulsive movements, and impaired function

G. Fine Motor Skills — strength and coordination of fine motor skills, e.g., writing, buttoning, use of scissors
1. Low muscle tone with sluggishness and impaired function
2. Fine motor activity is efficient, effective, and developmentally/age appropriate
3. Excessive muscle tone, muscle stiffness, poor coordination, and impaired function

8) SOMATIC AND SENSORY FUNCTIONS
A. Eating — maintaining dietary intake and nutrition
1. Insufficient diet interferes with nutrition
2. Maintains an appropriate diet
3. Excessive food intake, and/or unbalanced diet interfering with nutrition

B. Sexual/Gender Identity and Behavior — expressing age-appropriate sexual/gender identity and behavior
1. Denies or fails to recognize one's own sexual/gender identity, roles, or urges
2. Age-appropriate expression of one's own sexual/gender identity, roles, and urges
3. Overexpression of one's own sexual/gender identity, roles, or urges

C. Sensory reactivity — Response to sensory input (touch, taste, smell, sight, hearing, pain, and somatic)
1. Unresponsive to sensory input
2. Adaptively responds to sensory input
3. Hyperresponsive to sensory input, disrupting adaptive function

9) DAILY ROUTINES

A. Morning Routine — performing routines for starting the day at school

1. No evident routines present for starting the day
2. Appropriate, adaptive routines for starting the day
3. Rigid/inflexible routines that disrupt starting the day

B. Transitions — transitioning between activities or places

1. Cannot manage transitions, including routines for departure
2. Uses age-appropriate strategies and routines for transitions
3. Disruptive or rigid routines for transitions

C. Screen Time — managing screen-based activities (education, games, entertainment, social media)

1. Unwilling or unable to participate in screen-based activities
2. Age-appropriate use of screen-based activities
3. Social and/or academic functioning disrupted by screen-based activities

10) OTHER SPECIAL ISSUES

A. Toileting problems - toileting behaviors and routines

[ ] Present [ ] Not Present

B. Inadequate Self-Care/Hygiene - failure to appropriately care for one's own grooming, clothing, and general cleanliness

[ ] Present [ ] Not Present

C. Substance Use - using inappropriate substances/drugs (including alcohol, tobacco, cannabis, inhalants, biologicals, etc.)

[ ] Present [ ] Not Present

D. Trauma History - having directly or indirectly experienced traumatic events oneself or for others

[ ] Present [ ] Not Present

E. Distorted Perception of Reality - Inappropriate and/or inaccurate interpretation of events, experiences, and/or thoughts

[ ] Present [ ] Not present

F. Sleep Disturbance - Evidence of inadequate duration or quality of sleep, drowsiness, excessive yawning, persistent lack of energy/concentration/alertness

[ ] Present [ ] Not Present

G. Physically Harms Self - causing injury to self

[ ] Present [ ] Not Present

H. Physically or Emotionally Harms Others - causing physical or emotional harm to other persons or animals

[ ] Present [ ] Not present

I. Tics and other Atypical Movements are
[ ] Present [ ] Not Present

J. Bullying - Involved as Victim, Perpetrator, and/or Encouraging Observer
[ ] Present [ ] Not Present

Annex 6. Information form for teachers

Information Form for Participation in a Research Study

Learning Planet Institute

Principal Investigator: Ariel Lindner

Study Title: Learning through Iterative systems for Social-emotional Achievement (LISA)

Telephone: +33188328305

Email: ariel.lindner@cri-paris.org

Address: INSERM UMR1284, 8bis Rue Charles V, 75004 Paris, Learning Planet Institute (former CRI), 5th floor, office 5.08

You are being asked to participate in a research study. This form will give you the information you will need to understand why this study is being done and why you are being invited to participate. It will also describe what you will need to do to participate and any known inconveniences or discomforts that you may have while participating. We encourage you to take some time to think this over and ask questions now or at any other time.

PURPOSE

Educational processes and plans are often created with an “average” or, preferably, “typical” child in mind. This approach does not work well for many children. We seek to adapt educational processes to address each student’s needs and skills. To do this, we will use a structured approach to assist teachers in using their professional experience to develop classroom strategies based on each student’s individual needs. Using this process, we expect this to increase students’ sense of well-being while also helping them to learn better.

WHERE WILL THE STUDY TAKE PLACE AND HOW LONG WILL IT LAST?

The study will take place in multiple schools, including iféa, and will last until the end of the school year of 2023.

WHAT WILL I BE ASKED TO DO?

You will be asked to fill out one anonymous survey, participate in one group discussion, fill a questionnaire about each of your students twice, and to attend two teacher conferences that will be observed by the study staff.

The following information and data will be collected as part of the research project:

- Anonymous feedback survey about the questionnaire
- Focus group notes
- Your professional email address
- Parent’s email address
- Questionnaire responses
- Teachers’ meeting recordings
- Teacher’s meeting notes

DATA PROCESSING AND PRIVACY PROTECTION

Individual information and data will be stored privately, on the password and firewall-protected servers that are managed by the Learning Planet Institute. Meeting recordings will only be accessible

to the study staff under supervision of the principal investigator of this study. The recordings will be destroyed after a transcription of the recording has been performed. All names and other identifiable information mentioned in the transcription will be removed. All data will be deleted five years after the latest publication using this data. Before the destruction, you, the student or their legal guardians have the right to access, rectify, or delete their data, and to oppose or limit the data treatment. To request any of the aforementioned actions please contact the study coordinator, Kseniia Konishcheva, MSc, by phone +33769016161, or by email, kseniia.konishcheva@cri-paris.org.

All information and data about participants shared outside the study staff will be de-identified by default. This includes not sharing specific details such that an individual's identity may be inferred.

If you feel, after contacting us, that your rights to Information Technology and Liberties are not respected, you can send a complaint to the CNIL (Commission Nationale Informatique et Libertés www.cnil.fr).

WHAT IF I HAVE QUESTIONS?

Take as long as you like before you make a decision. We will be happy to answer any question you have about this study. If you have further questions about this project or if you have a research-related problem, you may contact the study coordinator, Kseniia Konishcheva, MSc, by phone +33769016161, or by email, kseniia.konishcheva@cri-paris.org.

CAN I STOP BEING IN THE STUDY?

You do not have to participate in this study if you do not want to. If you agree to be in the study but later change your mind, you may drop out at any time. There are no penalties or negative consequences of any kind if you decide that you do not want to participate. To withdraw and request deletion of information you provided, please contact the study coordinator, Kseniia Konishcheva, MSc, by phone +33769016161, or by email, kseniia.konishcheva@cri-paris.org.

Participation provides consent. This project has received a favorable opinion from INSERM Ethical Evaluation Committee (IRB00003888) on 12/04/2022 (n°22-897).

Annex 7. Information form for student's parents

Information Form for Participation in a Research Study

Learning Planet Institute

Principal Investigator: Ariel Lindner

Study Title: Learning through Iterative systems for Social-emotional Achievement (LISA)

Telephone: +33188328305

Email: ariel.lindner@cri-paris.org

Address: INSERM UMR1284, 8bis Rue Charles V, 75004 Paris, Learning Planet Institute (former CRI), 5th floor, office 5.08

You are being asked to participate in a research study. This form will give you the information you will need to understand why this study is being done and why you are being invited to participate. It will also describe what you will need to do to participate and any known inconveniences or discomforts that you may have while participating. We encourage you to take some time to think this over and ask questions now or at any other time.

PURPOSE

Educational processes and plans are often created with an “average” or, preferably, “typical” child in mind. This approach does not work well for many children. We seek to adapt educational processes to address each student’s needs and skills. To do this, we will use a structured approach to assist teachers in using their professional experience to develop classroom strategies based on each student’s individual needs. Using this process, we expect this to increase students’ sense of well-being while also helping them to learn better.

WHERE WILL THE STUDY TAKE PLACE AND HOW LONG WILL IT LAST?

The study will take place in multiple schools, including iféa, and will last until the end of the school year of 2023.

WHAT WILL I BE ASKED TO DO?

We will ask you to complete a questionnaire about your child in English. The questionnaire covers 62 behaviors that often impact students’ learning. The questionnaire’s purpose is to help teachers develop strategies to improve each student’s learning experience. Your responses to the questionnaire will be used to measure how reliable the questionnaire is at assessing the relevant behaviors. The questionnaire helps understand a child’s strengths and needs, for developing appropriate educational support.

The following information and data will be collected as part of the research project:

- Your name and relationship to the student
- The student’s school and teachers
- Questionnaire responses about your child
- Your email address

DATA PROCESSING AND PRIVACY PROTECTION

At the time of data collection, we will code the relationship between each respondent (parent and teacher) and the student being assessed, and remove all identifying information (name, email) for the students and the respondents.

Individual information and data will be stored privately, on the password and firewall-protected servers that are managed by the Learning Planet Institute.

All information and data about participants shared outside the study staff will be de-identified by default. This includes not sharing specific details such that an individual's identity may be inferred.

All data will be deleted five years after the latest publication using this data. Before the destruction, you have the right to access, rectify, or delete all data, and to oppose or limit the data treatment.

If you feel, after contacting us, that your rights to Information Technology and Liberties are not respected, you can send a complaint to the CNIL (Commission Nationale Informatique et Libertés www.cnil.fr).

WHAT IF I HAVE QUESTIONS?

Take as long as you like before you make a decision. We will be happy to answer any question you have about this study. If you have further questions about this project or if you have a research-related problem, you may contact the study coordinator, Kseniia Konishcheva, MSc, by phone +33769016161, or by email, kseniia.konishcheva@cri-paris.org.

CAN I STOP BEING IN THE STUDY?

You or your child do not have to participate in this study if you do not so desire. If you agree to be in the study but later change your mind, you may drop out at any time. There are no penalties or negative consequences of any kind if you decide that you do not want to participate. To withdraw and request deletion of information you provided, please contact the study coordinator, Kseniia Konishcheva, MSc, by phone +33769016161, or by email, kseniia.konishcheva@cri-paris.org.

Participation provides consent. This project has received a favorable opinion from INSERM Ethical Evaluation Committee (IRB00003888) on 12/04/2022 (n°22-897).

Annex 8. Information form for students

Formulaire d'information pour la participation à une étude de recherche

Institut Learning Planet, INSERM, Université de Paris Cité

Dans le cadre de cette étude, vos enseignants et vos parents rempliront un questionnaire à votre sujet et en discuteront ensuite pour savoir comment vous aider au mieux à l'école.

Nous vous demandons de participer à une étude de recherche car nous voulons en savoir plus sur la façon dont les enseignants et les parents peuvent aider leurs élèves à réussir à l'école.

Si vous acceptez de participer à cette étude, vous n'aurez rien à faire. Cela ne vous prendra pas de temps et ne vous causera aucun désagrément.

Veuillez en parler avec vos parents avant de décider de participer ou non. Nous demanderons également à vos parents de donner leur autorisation pour que vous participiez à cette étude. Mais même si vos parents disent "oui", si vous ne voulez pas participer à cette étude, vous n'êtes pas obligé de le faire. N'oubliez pas que c'est vous qui décidez de participer à cette étude et que cela ne dérangera personne si vous ne voulez pas le faire. Si vous participez à l'étude et que vous changez d'avis par la suite, vous pouvez arrêter de participer à tout moment. Personne ne sera fâché ou contrarié contre vous.

Vous pouvez poser toutes les questions que vous souhaitez sur l'étude. Si, plus tard, vous avez une question à laquelle vous n'avez pas pensé maintenant, vous pouvez nous appeler au +33769016161.

Pour arrêter de participer à l'étude, veuillez contacter la coordinatrice de l'étude, Kseniia Konishcheva, par téléphone au +33769016161, ou par e-mail, kseniia.konishcheva@cri-paris.org.