UMEÅ UNIVERSITY

Adolescent mental health

Time trends and validity of self-report measures

Ida Blomqvist

Department of Clinical Science
Child- and Adolescent Psychiatry
Umeå 2021

Dissertation for PhD
ISBN print: 978-91-7855-674-8
ISBN PDF: 978-91-7855-675-5
ISSN: 0346-6612 New Series No 2153
Cover photo by Tomas Thelin
Electronic version available at: http://umu.diva-portal.org/
Printed by: CityPrint i Norr AB
Umeå, Sweden 2021

To Tomas, Juni, Stig and Maj

Fortsätt vara nyfiken! (Stay curious!)

Gun och Stig Blomqvist

Table of Contents

Abstract

Background: Studies of time trends of adolescent self-reported mental health suggest an increase of mental health symptoms globally. Unfortunately, several studies within the field have methodological problems, such as short time-period between measurements and different mental health measures over time. When estimating mental health through self-report measures, the measures need to be both valid and reliable. Reports from the Swedish National Board of Health and Welfare have shown that several self-report scales used in Child- and Adolescent Psychiatry lack validation in Swedish, and some are direct translations of adult self-report scales without proper age-adaption.

Aims: This thesis aims to add to previous knowledge regarding time trends of self-reported mental health among Swedish youth and to validate internationally used reliable self-report measures for use in Sweden.

Methods: In Study I, we investigated changes in self-reported mental health symptoms, both internalized and externalized, in two samples: The first sample in 1981 and the second in 2014, both samples including all grade 9 students of Luleå. The same composite self-report measures were used at both time points. In study II we translated and validated the Reynolds Adolescent Depression Scale second edition (RADS-2) with classical test theory. In study III, eight pediatric Patient-Reported Outcomes Measurement Information System (PROMIS®) item banks were translated to Swedish and culturally adapted using the Functional Assessment of Chronic Illness Therapy (FACIT) methodology. Study IV describes the item response theory (IRT) validation of two item banks, the PROMIS Pediatric Bank v2.0 – Anxiety and the PROMIS Pediatric Bank v2.0 - Depressive Symptoms, in a school- and Child- and Adolescent Psychiatry patient sample.

Results: Study I: There has been an increase in internalizing symptoms, especially among girls. Externalizing symptoms have decreased, especially among boys, and in 2014 compared to 1981; there is no significant difference between girls and boys. Study 2: The factor structure of the Swedish version of RADS-2 was confirmed and measurement invariance for sex and age-group. Reliability was acceptable to excellent for all subscales and the RADS-2 total scale. Concurrent, convergent, and discriminant validity was acceptable. Study III: All of the eight pediatric PROMIS item banks had translation issues to resolve. However, the translated and adapted versions were linguistically acceptable. Study IV: After removing a few items, the pediatric PROMIS item banks of anxiety and depressive symptoms showed good IRT fit statistics and no differential item functioning. A computer adaptive test (CAT) simulation supports the idea of the item banks to be appropriate to use with CAT.

Conclusion: This study supports the previous knowledge pointing to a rise in self-reported mental health, especially among girls. Valid and reliable diagnostic measures are needed in Child- and Adolescent Psychiatry. RADS-2 is an internationally established measure, and the Swedish version is now validated in a relatively large school sample. Item response theory has several advantages compared to classical test theory. We have translated eight PROMIS item banks to Swedish, and two of them, anxiety and depressive symptoms, have been validated with IRT in a school- and patient sample.

Abbreviations

χ^2	Chi-Square
BYI-A	Beck Youth Inventories of Emotional and Social Impairment Anger
BYI-D	Beck Youth Inventories of Emotional and Social Impairment Depression
CAP	Child and Adolescent Psychiatry
CAT	Computer-Adaptive Testing
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CI	Confidence Interval
CTT	Classical Test Theory
DIF	Differential Item Functioning
DSM	Diagnostic Statistical Manual of Mental Disorders
DWLSSS	Robust Diagonal Weighted Least Square
EFA	Exploratory Factor Analysis
FACIT	Functional Assessment of Chronic Illness Therapy
fMRI	functional Magnetic Resonance Imaging
FSS	Functional Somatic Symptoms
GDPR	General Data Protection Regulation
GLM	General Linear Model
GRM	Graded response model

I	Information
IRT	Item Response Theory
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
KMO	Kaiser-Meyer-Olkin factor adequacy
MDD	Major Depressive Disorder
MI	Measurement Invariance
NIH	National Institute of Health
RADS-2	Reynolds Adolescent Depression Scale second edition
RMSEA	Root Mean Square Error Approximation
PHO	PROMIS Health Organization
PROMIS	Patient-Reported Outcomes Measurement Information System
PTM	Power Threat Meaning Framework
RCT	Randomized Controlled
RDoC	Research Domain Criteria
SE	Standard Error
TLI	Tucker-Lewis Index
UPOP	*Ungdomars upplevelse av psykisk ohälsa – psykometriska egenskaper i nya svenska versioner av test* (Adolescents' experiences of mental illness – psychometric properties of new Swedish versions of test)
WHO	World Health Organization

WHO-5	World Health Organization Wellness Index
WLS	Weighted Least Square

List of papers

Blomqvist, I., Henje Blom, E., Hägglöf, B., & Hammarström, A. (2019). Increase of internalized mental health symptoms among adolescents during the last three decades. European Journal of Public Health. doi:10.1093/eurpub/ckz028

Blomqvist, I., Ekbäck, E., Dennhag, I., & Henje, E. (2021). Validation of the Swedish version of the Reynolds Adolescent Depression Scale second edition (RADS-2) in a normative sample. Nordic Journal of Psychiatry, 75(4), 292-300. doi:10.1080/08039488.2020.1850858

Blomqvist I, Chaplin JE, Nilsson E, Henje E, Dennhag I. Swedish translation and cross-cultural adaptation of eight pediatric item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS). Journal of patient-reported outcomes. 2021;5(1):80-.

Blomqvist, I., Chaplin, J.E., Henje, E., Dennhag, I. Item response theory validation of the Swedish pediatric PROMIS item banks of anxiety and depressive symptoms in clinical and community samples. Manuscript.

Populärvetenskaplig sammanfattning

Denna avhandling syftar till att öka vår kunskap gällande eventuella skillnader i självskattad psykisk ohälsa hos ungdomar i Sverige över tid och vidare att ta itu med bristen på åldersanpassade, och tillförlitliga självskattningsskalor på svenska.

Studier som undersökt tidstrender avseende psykisk hälsa hos unga tyder på en ökning av internaliserade symtom framförallt hos flickor. Kungliga Vetenskapsakademin utkom år 2010 med en systematisk litteraturgenomgång av kunskapsläget gällande utveckling av psykisk ohälsa hos ungdomar. Rapportens slutsats var att den psykiska ohälsan hos ungdomar ter sig öka men att det saknas tillförlitliga studier. Då de studier som finns har ett antal metodologiska problem, såsom olika demografi i de jämförda grupperna, olika självskattningsskalor vid de olika mätpunkterna eller för kort tid mellan mätningarna. För att tillförlitligt kunna mäta psykisk ohälsa hos ungdomar behövs validerade självskattningsskalor. En kartläggning som gjordes av Socialstyrelsen 2009 visade att det förekom en mängd skattningsskalor inom barn- och ungdomspsykiatrin i Sverige, men att ett flertal av dessa skattningsskalor inte validerats i svenska populationer. Vissa skattningsskalor var direkta översättningar av vuxenversionerna och saknade åldersanpassning. Reynolds Adolescent Depression Scale second edition (RADS-2) är en skattningsskala för depression hos unga som används mycket inom klinik och forskning internationellt men som ej varit översatt eller validerad på svenska.

Patent-Reported Outcomes Measurement Information system (PROMIS) är ett projekt initierat av amerikanska National Institute of Mental Health (NIMH), med målet att avancera och förbättra självskattningsskalor, bland annat genom att använda moderna psykometriska metoder såsom Item Response Theory (IRT). IRT är en matematisk modell som utgår från sannolikhet och bygger på antagandet att en skala mäter en underliggande latent förmåga. En respondents svar på olika frågor fångar den underliggande förmågan på ett kontinuum från lågt till högt. IRT vilar också på antagandet att frågor kan ha olika svårighetsgrad och att det därmed finns information i hur individer svarar utifrån svårighetsgrad på frågan. Skalor som är utvärderade med IRT går att använda tillsammans med computer adaptive testning (CAT), dvs datorstyrda självskattningsskalor. CAT är uppbyggt genom en algoritm som är förinställd som presenterar frågor till respondenten, frågorna kommer styras av hur respondenten svarat. Fördelen med CAT är att metoden kräver färre frågor jämfört med hur traditionella självskattningsskalor presenteras samtidigt som de är de precisa och valida.

I studie I som var en upprepad tvärsnittsstudie där åk 9 elever från Luleå kommun svarade på enkätfrågor, i två omgångar år 1981 och 2014. Det var hög svarsfrekvens vid båda mätningarna och vid de två tillfällena användes samma självskattningsskala. Självskattningsskalan som utvecklades 1981 har senare kopplats samman till olika kompositmått, med bland annat diagnossystemet DSM systemet som förlaga. Måtten mäter olika former av internaliserande och utåtagerande symtom. Analys av data visade signifikanta skillnader vid de två mättillfällena. Både flickor och pojkar skattade högre år 2014 än 1981 och flickor skattade högre än pojkar både 1981 och 2014. Ökningen i internaliserande symtom var större för flickor än pojkar dvs skillnaden mellan flickor och pojkar var större 2014 jämfört med 1981. För uppförandeproblem skattade pojkar högre än flickor 1981 men 2014 skattade både flickor och pojkar mindre utåtagerande symptom jämfört med 1981 och det var inte längre någon signifikant skillnad mellan flickor och pojkar.

I studie 2 har vi översatt RADS-2 till svenska och validerat den hos skolungdomar. RADS-2 används för att mäta allvarlighetsgrad av depression och används i stor utsträckning internationellt men är ännu inte validerad på svenska. RADS-2 visade sig samvariera så som önskat med andra etablerade skattningsskalor för depression såsom Becks ungdomsskalor för nedstämdhet. Faktorstrukturen blev också bekräftad i en faktoranalys och RADS-2 hade höga Cronbach's alfavärden, vilket tyder på god interkonsistens, ett reliabilitets mått som används inom klassisk testteori. Sammantaget ter sig RADS-2 fungera väl för skattning av depressiva symptom hos svenska skolungdomar.

I studie 3 var syftet att översätta och utvärdera översättningen av åtta PROMIS självskattningsskalor till svenska. Arbetet följde Functional Assessment of Chronic Illness Therapy (FACIT) översättningsmetod, med en modifikation, nämligen användandet av granskningsgrupper. Granskningsgrupperna bestod av personer som var professionellt verksamma inom bland annat lingvistik, psykometri och barn- och ungdomspsykiatri. Både forskare och kliniker gick genom och diskuterade alla åtta PROMIS-skalorna. Förutom detta steg så ingick också översättning från originalspråk till svenska, återöversättning till originalspråket engelska samt kognitiva intervjuer. De kognitiva intervjuerna genomfördes med barn i åldrarna 8–17 år med syftet att utvärdera hur de uppfattade PROMIS självskattningsskalorna. När alla delar av processen utförts blev slutsatsen att PROMIS självskattningsskalorna är redo för fortsatt validering på svenska.

I Studie IV har PROMIS självskattningsskalor för ångest- och depressionssymtom utvärderats med IRT och CAT-simulering hos en grupp ungdomar från både skolor och Barn- och Ungdomspsykiatrin (BUP). Användandet av IRT vid validering har många fördelar jämfört med traditionella

metoder för validering (klassisk testteori) t.ex. så erhålls för varje fråga en uppfattning om dess svårighetsgrad i att mäta den underliggande förmågan. Det går också att utvärdera så att varje fråga uppfattas lika, dvs inte mäter olika på grund av grupptillhörighet såsom kön och åldersgrupp. Sammanfattningsvis så har självskattningsskalorna visat sig fungera bra och kan med fördel användas med CAT i fortsättningen.

Våra resultat tyder på att psykisk ohälsa ökat hos ungdomar framför allt flickor under de sista trettio åren. I ett försök att teoretisera kring orsakerna till denna utveckling använde vi Urie Bronfenbrenners ekologiska systemteori. Denna postulerar att barn och ungas mående och utveckling påverkas av flera omgivande system och att dessa interagerar med varandra med en potentiellt ackumulerande effekt på barnet. Larsson et al. från Göteborgs Universitet skriver i boken ”Transformations of the Swedish welfare state: from social engineering to governance?” att vi under de senaste årtiondena i Sverige gått från ett välfärdssamhälle till ett samhälle med mer neoliberalistiska strömningar. Därmed har självkontroll och förmågan att vara ”sin egen lyckas smed” blivit alltmer viktigt, vilket ökar stressen på unga. Flera rapporter från Skolverket tyder också på att socioekonomiska skillnader påverkar skolresultaten, exempelvis läsförmåga i åk 4.

I arbetet med denna avhandling har vikten av tillförlitliga självskattningsskalor blivit tydlig. Det är lätt hänt att resultaten från en skattningsskala tolkas som en absolut sanning. Det är viktigt att vara försiktig med tolkningar och slutsatser av skattningsskalor särskilt när det i utvärderingsprocessen ofta finns metodologiska begränsningar. Dessutom finns det anledning att fundera kring den diagnostiska validitet för de diagnossystem som används som mall för de flesta skattningsskalor, särskilt för barn och ungdomar. Det går alltid att tolka resultat från en skattningsskala kvalitativt och använda skattningsskalor som ett sätt att öppna upp för ytterligare frågor och djupare förståelse för respondenten. Om resultaten ska tolkas kvantitativt så är det t.ex. viktigt att skalan är validerad för svenska ungdomar, dvs. i ett stickprov som utgör liknande population som den som respondenten tillhör. Min slutsats är att PROMIS och användandet av IRT ger hopp för att öka kvaliteten av skattningsskalor och därmed en säkrare psykiatrisk diagnostik och behandling av unga.

1. Introduction

This thesis aims to increase our knowledge of the variation in self-reported mental health symptoms in adolescents in Sweden over time and address the lack of Swedish-language, age-adapted, reliable, and valid self-report scales.

There have been several studies on time trends in mental health. However, these have had various methodological problems, such as different demographics in the compared groups, different mental health measures over time, or short time between measurements. Furthermore, many self-report measures used in Child and Adolescent Psychiatry clinics (CAP) in Sweden have not been validated in Swedish. Some are direct translations of adult measures without age adaption. The Reynolds Adolescent Depression Scale second edition (RADS-2) measures symptom severity of depression and is extensively used internationally but is not yet validated in Swedish. Therefore, we set out to validate RADS-2 in Swedish adolescents. In the process of validating RADS-2 with the most common psychometric method, Classical Test Theory (CTT), the disadvantages with CTT became evident, and the need to further explore alternative psychometric methods, such as Item Response Theory (IRT), became apparent. The US National Institute of Health has funded the Patient-Reported Outcomes Measurement Information System (PROMIS) project. PROMIS aims to advance measurement scales through the use of IRT, and the last part of this thesis focuses on the translation of several pediatric PROMIS item banks to Swedish. In addition, we provide an item response theory validation of the item banks of anxiety and depressive symptoms in a school and CAP-clinic patient samples.

Here, a short review of the current state of the art in the field is given to provide a rationale for the work.

2. Background

2.1 Time trends in adolescent mental health

Studies of time trends, including a meta-analysis of internalized mental health symptoms among young people from the 1980s to the early 2000s, suggest an increase of mental health symptoms globally, especially internalizing symptoms in adolescent girls. At the same time, the development among boys is less clear (1, 2). After the early 2000s, a more stable increasing trend has been noted (3). A limitation of the meta-analysis was the relatively low amount of studies included because there were not many studies of sufficient quality. Studies of externalized symptoms in the general population also indicate an increasing trend from the 1970s to the 1990s (2, 4). After that, the trend levels off. A review covering 1985–2011 indicates stable levels of externalized problems in Swedish youth, with only a few of the included studies reporting increasing levels of these symptoms for girls (1).

Unfortunately, several studies in the field have methodological problems, such as comparing groups with different demographics, using different mental health measures over time, a short time between measurements, and inconsistent reporting on gender differences (1, 5). Furthermore, most studies lack comparable measures of mental health over time. Repeated cross-sectional investigations using the same mental health measures in geographically and socially comparable groups are needed (6-8).

In Sweden, data from the World Health Organization (WHO) report "Health Behavior among School-aged Children" show an increased prevalence of headache, stomach ache, backache, and dizziness among 15-year-olds, predominantly girls, from the mid-1980s to today (9). Psychiatric symptoms such as insomnia, depression, irritability, and nervousness have also increased in 15-year-olds, especially among girls (9). The Royal Swedish Academy's Health Committee's state-of-science conference in 2010 reported that there seems to be an increase in depressive and anxiety symptoms in children and adolescents over the last decades. However, they concluded that the knowledge base is thin, with too few studies with insufficient quality, and therefore more research is needed for conclusive results (10, 11).

There seems to be a congruence between self-rated internalized symptoms in particular and clinical diagnoses of anxiety syndromes and major depressive disorder (MDD) (12, 13). At present, MDD accounts for a major part of the global disease burden in adolescents and young adults (14). The global projections of the WHO predict that unipolar depression will be the leading cause of the global burden of disease by 2030 (15). The consequences of this trend could be severe since adolescent depression increases the risk of suicide and is associated with considerable present and future morbidity (16). Early-onset MDD increases the risk of recurrent depressive episodes in adulthood and is related to a more severe course of the disease (17-19).

2.2 Measures of mental symptoms in adolescents

In child and adolescent psychiatry, self-report questionnaires are used to assess symptoms like depression and anxiety. The scales are used as proxies for variables that cannot be directly observed, such as the subjective experience of a patient's inner depressive feelings, cognitions, and symptoms. By summing the rating of each item, a total score can be calculated as an indication of symptom severity. Because no objective biomarkers have been identified that can aid the diagnostic process of psychiatric disorders to date, self-reported measures and clinical interviews constitute the primary diagnostic support for the clinician.

At present, various types of measures are being used to estimate adolescent mental health. For example, depressive symptom severity can be measured by clinical assessment and self-reporting or parent/teacher-report instruments. In general, for internalizing symptoms in adolescents, self-reported measures are considered to give better information than parent or teacher reports (20, 21). According to reports from the Swedish National Board of Health and Welfare, an array of different instruments is in use both for screening purposes (22) and for clinical use, such as measurement of the treatment outcome (23). About half of the instruments have not been investigated regarding psychometric properties in the Swedish versions (23, 24). Some instruments are direct translations of adult depression self-rating scales and do not consider the difference of depression symptomatology between adult and adolescent depression. In addition, there are no Swedish instruments for adolescent depression symptom severity rating

(which are compatible with the Diagnostic Statistical Manual of Mental Disorders (DSM) system), nor any instruments that provide subscales measuring different aspects of MDD (22-24).

2.3 The Patient-Reported Outcomes Measurement Information Systems (PROMIS)

The Patient-Reported Outcomes Measurement Information Systems (PROMIS) originated as a National Institute of Health (NIH) Roadmap Initiative. The purpose was to address major gaps in biomedical research that "no single NIH institute could tackle alone, but which the agency as a whole [could] address to make the biggest impact possible on the progress of medical research" (25). In the process, a multi-center cooperative group was founded to create:

"…new paradigms for how clinical research information is collected, used, and reported. PROMIS addressed a need in the clinical research community for a rigorously tested patient reported outcome (PRO) measurement tool, that uses recent advances in information technology, psychometrics, and qualitative, cognitive, and health survey research to measure PROs such as pain, fatigue, physical functioning, emotional distress, and social role participation that have a major impact on quality-of-life across a variety of chronic diseases." (26)

PROMIS has several advantages over traditional questionnaires, such as the utilization of item response theory (IRT) in the validation process. IRT validated measures or, in other words, item banks, are set of items calibrated to the same underlying construct. Item banks can, in turn, be utilized with Computer Adaptive Testing (CAT), the advantages of which will be discussed in the next section. PROMIS has been shown to have higher reliability, validity, and better sensitivity to change than most legacy health measures (27-31), and responsiveness has also been shown to be better than traditional questionnaires (30, 32, 33). PROMIS has now been used in several studies in children, adolescents, and adults with varying mental and physical conditions (34-38). Several PROMIS item banks have been translated and used in various countries such as Brazil, the Netherlands, and China during recent years (39-44).

2.3.1 Translation of PROMIS item banks

Various methods are available for translation and cultural adaption of items, such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) (45) and the Functional Assessment of Chronic Illness Therapy (FACIT) (46) methodologies. Both have a multistep approach, e.g., including translation, back translation, reconciliation, cognitive debriefing. However, a review evaluating translation methods recommended a multistep process and noted that different methods can be used while still obtaining similar results (47). Therefore, the research team conducting the translation is granted the freedom to choose methodology according to logistics and preferences. Differential item functioning (DIF) tests whether an item is perceived differently, e.g., due to potential group prerequisites, such as language, gender, socioeconomic situation. DIF tests between languages are a possible method for further evaluating the translations of the items between languages (45).

2.4 Overview of methods for validating self-report measures

2.4.1 Classic Test Theory (CTT)

Classic Test Theory (CTT) is sometimes called 'true score theory .' It is based on the understanding that the sum of the score of any given questionnaire has two components (48, 49).

Observed score = True score + Error

For any given response, the observed score will reflect the 'true' score but also (inevitably) include an error score. Whether the errors for each question are negative or positive is assumed to be random and independent from each other. Therefore, their mean is assumed to be zero, and when all errors are combined, they should cancel each other out (48). However, many circumstances can affect the error, such as the setting in which the test is taken. For example, a noisy setting or place with lots of disturbances might generate a different outcome than

a quiet place without disturbances. Also, the state of the respondent might influence the results if the respondent is troubled or has not slept well, or if they, on the contrary, are feeling focused and motivated. These circumstances are assumed to be the same for all respondents and should cancel each other out in theory (48).

Standard error of measurement describes the expected fluctuation of a test due to error. Statistically, if two-test takers have a different score, but the confidence interval of the standard error overlaps, the difference between the respondents is not considered significant (at a predefined significant level such as 0.05). In CTT, one standard error of measurement is obtained for the whole scale. The validation is sample-dependent, i.e., the parameter estimates of the scale will depend on the sample (57).

When testing the psychometric properties of a psychiatric self-report questionnaire, conclusions are drawn about the test's reliability and validity. Reliability refers to whether a measure is stable and repeatable for different persons, occasions, conditions, and time points. Here, we will focus on two different types of reliability measures, namely test-retest and internal consistency. Test-retest is measured by the correlation between the two different measurement points (50). In contrast, internal consistency measures the average intercorrelations among the single items often tested using Cronbach's alpha (51, 52) and is commonly used for establishing reliability in CTT.

Validity can be measured and conceptualized in a few different ways. In this thesis, the term 'concurrent validity' will be used to refer to the potential overlap between a new measure and a well-established, previously validated measure. 'Convergent validity' will be used to refer to how well a total score correlates with a measure of associated constructs. 'Discriminant validity' will be used to measure how well any specific instrument differentiates the construct it measures from other constructs (53, 54). Each of these validity measures will be tested by correlations with measures of the same (concurrent), similar (convergent), or different (discriminant) constructs.

2.4.1.1 Disadvantages of CTT

CTT-validated scales only yield one standard error of measurement (SEM or standard error SE), that is, one SE for the full scale (55). The SE will be regardless of the respondents' latent ability estimate, regardless of the respondents' scored low, medium, or high on the scale, unlike scales validated with item response theory (IRT) (56, 57). IRT validated scales obtain an SE at a point estimate of the latent ability. Hence, the standard error of measurement will depend on the estimate of the respondents' latent ability (57). Therefore, CTT validated scales are less accurate than scales validated with IRT.

CTT-validated scales enumerate all item scores in the measure, with the same weight for all items, into a total score. Summated scores can be problematic because a symptom (e.g., crying) may not be as critical relative to others (e.g., suicidal thoughts) in measuring a specific trait (e.g., depression) but yet will hold the same level of importance. Ideally, core symptoms of the underlying trait, e.g., depression, such as suicidal thoughts, should be influencing the overall score more than symptoms less critical such as crying. The weighting of the items is not possible with CTT-scales; all items will have equal weight on the total score. Furthermore, for CTT validated scales, the more items in the scale, the higher reliability (48, 57). Many items in the scale add a risk of having items that are more of a noise than actual underlying core traits (of the measured construct) and that the sum of these incorrectly point to a diagnosis such as depression.

CTT measures are sample dependent, which means that, in theory, a new validation should take place for every new sample. The results of any given validation cannot be generalized to a new sample (56, 58, 59). Therefore, it is troublesome if the scale is used to draw conclusions on patient diagnosis or severity of a disease if the patient does not adhere to the sample the measure is validated in, such as the use of scales in Sweden not validated in a Swedish context.

A disadvantage of CTT is that there is no robust way of evaluating group differences at an individual item level. In other words, we cannot assess whether scale items are measurement invariant. Therefore, systematic bias between

groups may exist, but they will not be identified, which may have significant implications in psychological research.

Another disadvantage with CTT is its inability to control for systematic errors (56) such as factors within the person, e.g., fatigue or inattention. This, in turn, has implications for the interpretation of raw scores. If the error term is not randomized, it will affect the observed score without the ability to know how much influence the error term has on the observed score.

Furthermore, the underlying assumption in classical test theory that the true score is normally distributed is also problematic since most psychological measures are not (56). This can lead to biased parameter estimates.

Whether Cronbach's alpha can be used with ordinal data is debated among researchers (60, 61). Cronbach's alpha assumes tau-equivalence, which is the idea that the same factor loading applies for all test items in a scale in a factorial model. This assumption is rarely met for most scales used in a psychiatric context. Cronbach's alpha also assumes that variables are continuous rather than ordinal (most variables in psychological measures are ordinal) and that there is no error variance. Finally, Cronbach's alpha assumes that the measure is only measuring one dimension. Therefore, for scales with several dimensions with a total score that sums the dimensions, such as the Strengths and Difficulties Questionnaire (SDQ) (62), Cronbach's is not appropriate to measure reliability for the total scale.

Another measure called ordinal alpha is conceptually similar to Cronbach's alpha but does not demand tau-equivalence. However, it also assumes that there is no variance between the error factors(63). Ordinal alpha has been proposed as a better measure than Cronbach's alpha for ordinal data; however, at least one researcher has pointed out that ordinal alpha is more of a theoretical measure and should be interpreted only as such (64). There are other types of theoretical (but rarely used) alternatives, such as Coefficient H (65) and Greatest Lower Bound (60). However, it is hard to find a reliability measure in which the assumptions are not violated by common factors of psychological measures, e.g., ordinal variables, skewed data, or tau-inequivalence.

2.4.2 Item Response Theory

Item response theory (IRT) is a conceptually different method than CTT. IRT uses a more probabilistic approach as opposed to CTT, which uses a more correlational approach. IRT rests on the assumption that a scale measures a latent construct and that the scale captures the respondent's ability on a continuum from low to high on that construct. In other words, it is an estimate of a person's underlying ability on that latent construct in question, and the probability of the endorsement of any single item rests upon that underlying ability. Consequently, the items on the scale capture a specific ability level of that construct. Items can also have different levels of difficulty, and answers about items are therefore differentially informative. E.g., different levels of information about shoulder function are obtained by the items "Can you touch your earlobe?" and "Can you raise your arm above shoulder level?" The respondent's answers to each item allow conclusions to be drawn about the respondent's ability. The IRT model will provide an information function that gives knowledge about the items' specific ability/difficulty level range. Therefore, it is possible to conclude the respondents' underlying abilities of the measured construct by how they answer the items. The item difficulty and a persons' latent ability are estimated on the same continuum.

The following assumptions are tested before applying an IRT model: unidimensionality, local independence, and monotonicity. Unidimensionality describes whether the scale measures a single construct and can, for example, be shown in a factor analytic framework (through exploratory factor analysis or confirmatory factor analysis) by all items loading highly on the same latent construct and with fit indices with-in predefined boundaries (31). Local independence assumes no significant association between item responses after controlling for the dominant factor. Thus, the dominant factor is the only thing that influences a person's response to an item. Items that risk not having local independence are, for example, items that are too similar, like "I feel lonely" and "I feel alone." Monotonicity is the probability of selecting an item response that corresponds with the trait levels that are being assessed. For example, monotonicity assumes that a person with a high level of depression will select an item response that will adequately reflect that level of depression.

After unidimensionality, local independence, and monotonicity has been established, an item response model can be applied. There are several different item response models; in this thesis, the graded response model (GRM) will be used. The GRM model is suitable for polytomous response data with ordered categories and is recommended for PROMIS item banks (31). In a graded response model, threshold values are estimated, e.g., on which ability level respondents will change from scoring 'never' to 'almost never' 'almost never' to 'sometime,' 'sometimes' to 'often,' and so on. Furthermore, thetas (θ) are obtained. Thetas are the underlying ability of the respondent on the constructs measured. Figure 1 shows an example of a graph of the changing probabilities, using theta (θ) as the measure of the underlying capability of the trait measured. The y-axis is the probability of a given theta P(θ), and the x-axis is the theta (θ) level. In the example shown, namely answers to PROD7 (the PROMIS depressive symptoms item bank, item 7), the threshold values of changes from one answer to the next (i.e., where the lines cross in the graph) are at 0.10, 0.67, 1.37, and 1.81 (62) on the x-axis.

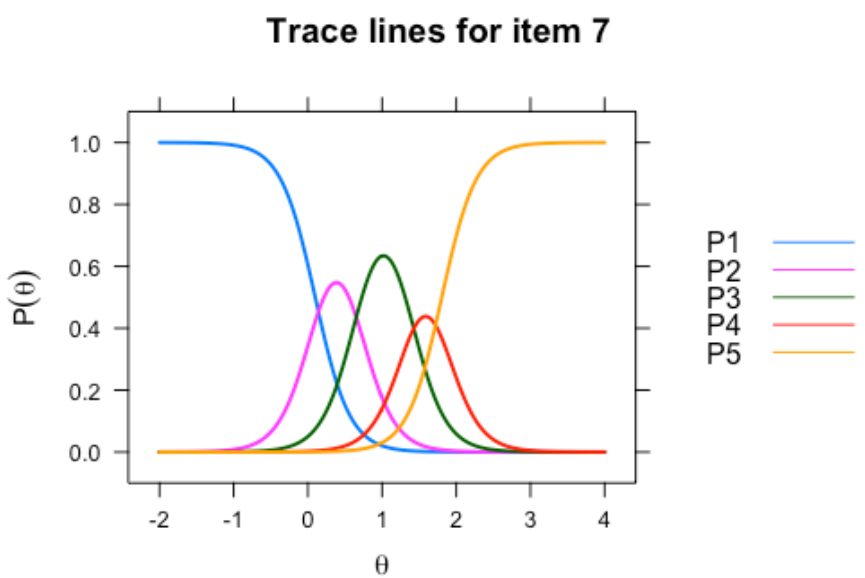

Figure 1. Trace lines for the pediatric PROMIS depressive symptoms item bank item 7. Blue line = never, pink line = almost never, green line = sometimes, red line = often, yellow line = almost always.

For every item calibrated with IRT, a standard error is obtained, and the individual standard error can be summed together, giving the whole item bank standard error. The standard error is inversely related to test information (I), as given by the equation (57):

$$SE(\theta) = 1/\sqrt{I(\theta)}$$

SE (θ)= standard error given theta

I(θ) = Information given theta

θ = theta (i.e., the underlying ability of respondent of the construct measured)

The plotting of the thetas and standard errors can give a graphical overview of which ability levels an item measures and its reliability. Reliability in the IRT framework can be assessed by the equation below (57):

$$r = 1 - SE^2$$

r = reliability

SE = standard error

For example, a standard error of 0.30 equals a reliability of 0.91 (that is, $1 - 0.30^2 = 0.91$) (57).

Figure 2 presents the test standard errors, and figure 3 shows the test information. As shown in Figures 2 and 3, the standard error is lowest (in conjunction with the highest information and, in turn, highest reliability), ranging between 0 to 2 on theta levels. Thetas between 0 to 2 indicate higher levels of reliability and progressively changes as it spreads to both ends of the ability spectrum (which indicates lower levels of reliability).

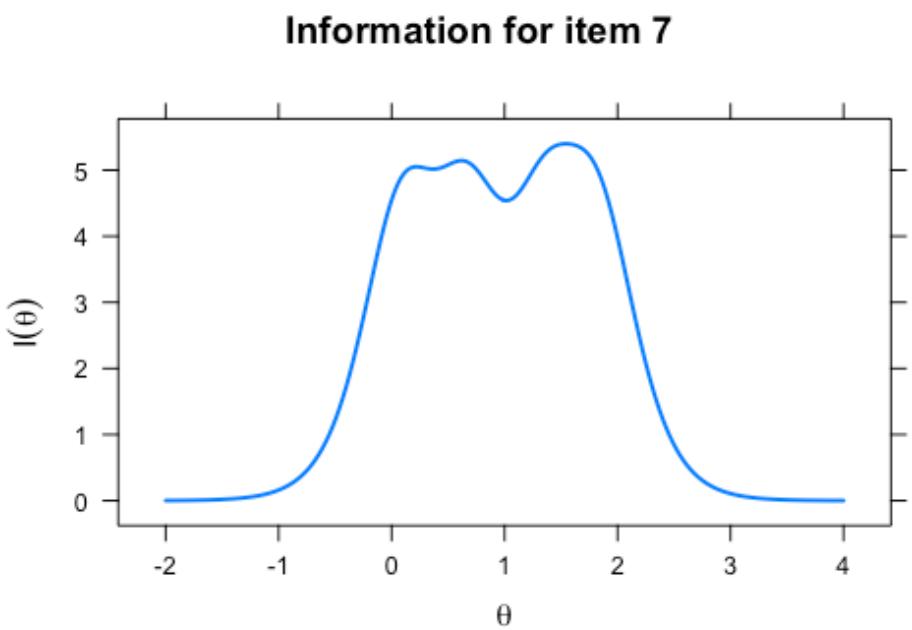

Figure 2. Test information and reliability for the PROMIS depressive symptoms item 7. $I(\theta)$ = Information given theta; θ= theta/ability level of the item.

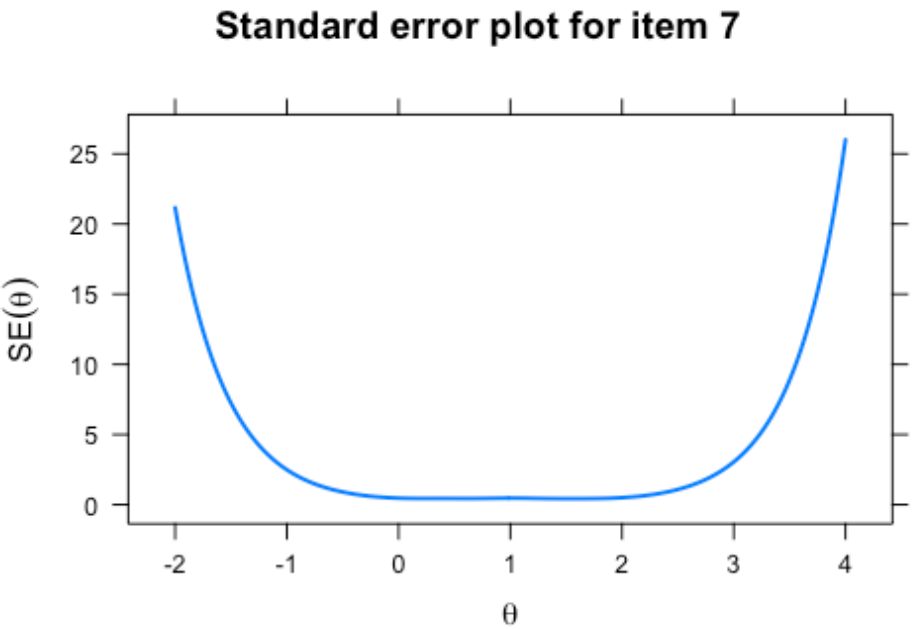

Figure 3. The standard error for the PROMIS depressive item 7. SE (θ) = standard error of the item given theta/ability; θ= theta/ability level of the item.

Figure 4 presents the test standard error, and figure 5 shows the test information for the full item bank. The figures can be interpreted similarly, as Figures 2 and 3. The standard error is lowest (in conjunction with the highest information and, in turn, highest reliability), ranging between approximately -3.5 to 4 on theta levels for the full item bank. Thetas between -3.5 to 4 indicate higher levels of reliability and progressively changes as it spreads to both ends of the ability spectrum (which indicates lower levels of reliability).

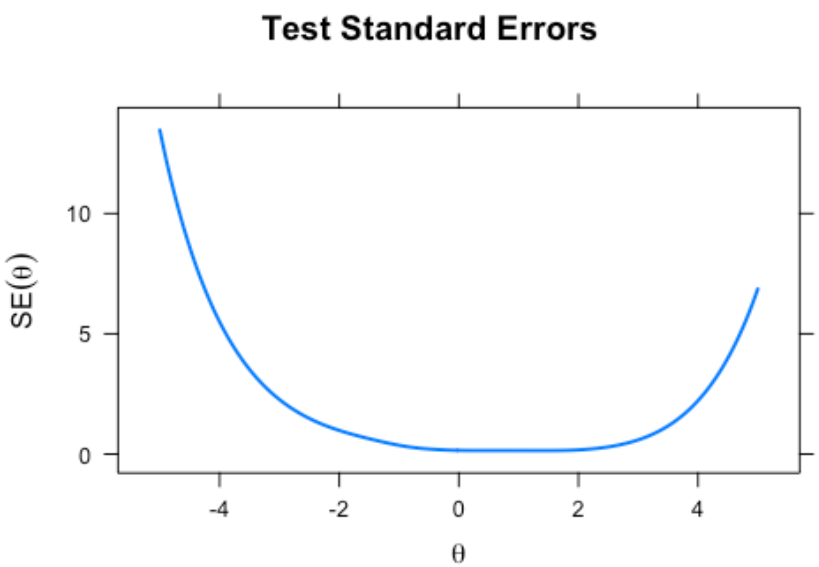

Figure 4. The standard error for the total PROMIS depressive symptoms item bank. SE (θ) = standard error of the item given theta/ability; θ= theta/ability level of the item.

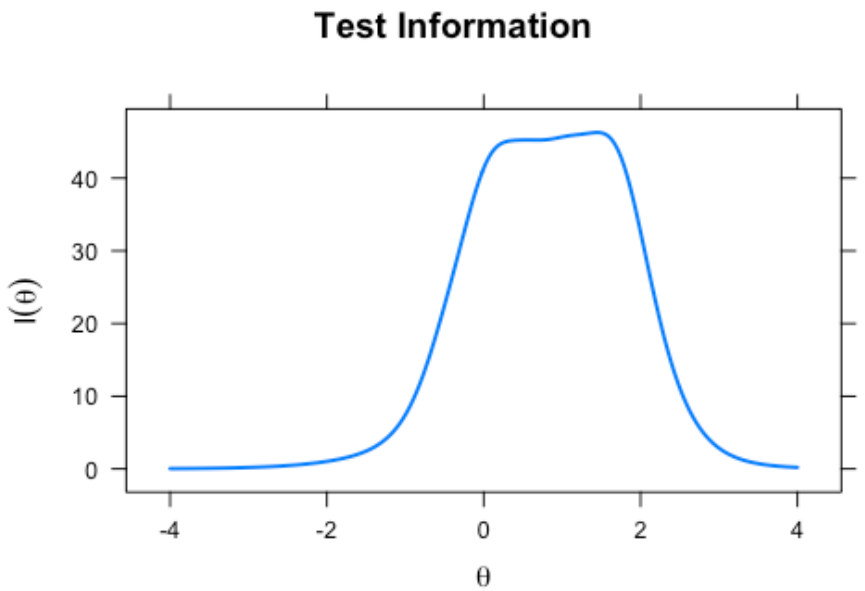

Figure 5. Information for the total PROMIS depressive symptoms item bank. $I(\theta)$ = Information given theta; θ= theta/ability level of the item.

2.4.2.1 Computer adaptive testing (CAT)

Computer adaptive testing (CAT) is a modern technological approach to present questionnaire items. With CAT, the items administered uniquely adapt to each respondent based on the attempts of current questions in the same test session. In other words, this method iteratively selects the following item and generally reduces the number of items needed to be presented and, consequently, the respondent's burden in having to answer all items in the bank while maintaining high levels of accuracy in estimating the respondents' latent ability.

The item parameters of the item bank used in a CAT are derived from an IRT model. An algorithm with a predefined stopping criterion will constrain the number of items being administered—a maximum of 12 items, or a predefined standard error of 0.3 are examples of stopping rules. The starting item is generally one with a medium difficulty level relative to the targeted population in order to acquire a first general idea of the respondents' level of the underlying construct.

The next item will vary depending on how the respondents answer the preceding/previous item, i.e., low, medium, or high difficulty. Therefore, the preceding items will be centered around the estimate of the respondents' underlying ability to get the best possible measurement preciseness. No more items are administered when the stopping rule criteria have been met. This thesis uses the Maximum Fisher Information (MFI) for the next item selection in the CAT (66, 67). This is one of the most popular selection methods that select the next item by maximizing the test information function.

CAT is a way to have a respondent answer fewer items but still maintain a low measurement error (although the lowest measurement error will almost always be achieved by answering the total item bank) (68). CAT can be used for screening purposes at Child- and Adolescent Psychiatry (CAP) clinics. Several item banks can be administered while still keeping the total amount of items low. Most scales currently in use in CAP have a large number of items, for example, Becks Depression questionnaires and Becks Anger questionnaire with 20 items each (the more items in the scale, the higher the reliability, when validated with CTT). In contrast, using a CAT item bank may produce a valid result after only 4 or 5 items. Thus, it is possible to administer several different item banks (e.g., fatigue, physical function, anxiety, depressive symptoms, or anger) without administering more than 20 items. However, while CAT may be efficient in item administration, the SE calculation changes depending on the position of the person's latent ability. Consequently, respondents on the ends (lower or higher) of the ability spectrum may be given more items. Furthermore, CAT can only be used on a computer and not via paper and pencil. Thus, this could be one of the reasons that may have hindered clinics from adopting the modern approach to assessing people's psychological state.

3. Aims

3.1 Overall aims

The overall aim of this thesis is to address two major research gaps: first, the lack of population-based repeated-measures studies of mental health in adolescents; second, the lack of age-adapted, reliable, validated, and internationally comparable self-report scales to assess mental health in adolescents

3.2 Specific aims

The specific aims of the individual studies in this thesis were:

Study I: to measure possible changes of self-reported mental health symptoms in two samples of grade 9 students (about 15 years old), one group in 1981 and the second in 2014, in the same geographical area of Northern Sweden, regarding internalized symptoms and conduct problems.

Study II: to test the psychometric properties of the Swedish version of Reynolds Adolescent Depression Scale second edition (RADS-2) in a school sample.

Study III: to translate and culturally adapt eight pediatric PROMIS item banks for Swedish use.

Study IV: to validate the Swedish PROMIS pediatric item banks of anxiety and depressive symptoms in a school and patient sample.

4. Methodological considerations

4.1 Description of the samples

4.1.1 The Luleå sample in Study I

4.1.1.1 Participants

Two samples were obtained, the first in 1981 (n = 1083, 46.7% girls, response rate 99.7%) and the second in 2014 (n = 682, 49.6% girls, response rate 98.3%). The selection procedure was the same and included all pupils in their last year of compulsory school (most of them 16 years old) in the Swedish municipality of Luleå.

4.1.1.2 Study design and settings

This study was cross-sectional and compared two separate but geographically identical groups of 9th-grade students from 1981 and 2014, respectively. It was performed in a middle-sized industrial municipality, Luleå, in Northern Sweden. The municipality was representative for Sweden concerning sociodemographic factors and health status among young people (14).

4.1.2 The UPOP samples in Studies II, III, and IV

4.1.2.1 Participants, school sample Studies II and IV

Participants were recruited from four junior and high schools from different socioeconomic areas in northern Sweden. This convenience sample included students from different school programs such as natural science, social science, media, and the arts. Permission was granted from the principals at the schools, and the class teacher gave the students information about the study. Additional information was provided by research assistants. Eight hundred ninety-seven students were asked to participate in the study, and 637 (71%) of them agreed. The mean age was 15.73 (SD = 1.76).

Seventy percent of the participants lived with both parents, and eighty-eighty percent of the participants were born in Sweden. A Swedish socioeconomic classification system was used (69, 70) to estimate the socioeconomic status of the participants' households. The distribution of the socioeconomic classification of the participants' parents was as follows: 17.2% workers, 28.4% assistant and intermediate non-manual workers, 32.1% professionals, civil servants, and executives, 7.1% self-employed of various kinds, and 15.2% unknown.

4.1.2.2 Participants, cognitive debriefing in Study III

Cognitive debriefing interviews were carried out in a sample of eleven healthy children/adolescents between 8 and 17 years old (9 girls and 2 boys, mean age 14 years, median age 14 years). The cognitive debriefing sample was recruited within the social networks of the researchers, with the inclusion criteria being in the suitable age range and fluency in spoken and written Swedish. The only exclusion criteria were any psychiatric diagnoses.

4.1.2.3 Participants, clinical sample in Study IV

Participants in the clinical sample for study 4 were recruited from child and adolescent psychiatry (CAP) clinics in Northern Sweden. The patient sample was recruited through fliers at the CAP clinics and staff working at the CAP clinics.

4.1.3 Other general factors

In the samples recruited from schools and clinics, absolute inclusion criteria were age between 12 and 20. Exclusion criteria were non-fluency in written Swedish and inability to complete online or paper forms (e.g., severe dyslexia).

Due to law restrictions and the General Data Protection Regulation (GDPR) implemented during data collection, the online forum used for the students had to be replaced, and paper forms were used for the patient sample before a proper GDPR platform could be established.

4.2 Study I

4.2.1 Measurements

In this study, pre-constructed composite measures of mental health symptoms were used (66). The composite measures were constructed from single-item questions and inspired by the Youth Self-Report scale, subscales from the Strengths and Difficulties Questionnaire, and the DSM system's diagnostic symptom criteria for anxiety and depression disorders. The following composite measures were developed for internalizing problems: depressive symptoms, anxiety symptoms, and functional somatic symptoms (FSS). For externalized behavior, items were dichotomized into 'occurrence' and 'non-occurrence and then summarized into a measure of conduct problems. The validity of these composite measures has been tested and found to be acceptable (71).

4.2.2 Data analyses

Data were analyzed with descriptive statistics, and the Mann–Whitney U test was used to assess potential differences between demographic data. We used a general linear model (GLM) for parameter estimates and a two-way ANOVA to compare groups. In the analyses, each subscale (symptoms of anxiety, symptoms of depression, FSS, and conduct problems) served as the dependent variable, and gender (girls, boys), year (1981 and 2014), and parental socio-demographics served as the independent variables. The GLM was done in two applications: Model 1 analyzed the crude model, which included gender, year, and the interaction between gender and year, and Model 2 controlled for parents' sociodemographic factors (i.e., occupational classification, country of origin, (un)employment and living arrangements). Missing values were excluded from analyses. All analyses were performed using IBM SPSS 24.

4.3 Study II

4.3.1 Instruments used for validation

Reynolds Adolescent Depression Scale second edition (RADS-2) (54) consists of 30 brief self-statements on a 4-point scale ranging from 'Almost never' to 'Most of the time.' The items are divided into four subscales/dimensions (dysphoric mood, anhedonia/negative affect, negative self-evaluation, and somatic complaints). The anhedonia/negative affect items are formulated with positive questions such as "I feel happy" and thus are reverse coded. Higher scores indicate higher symptom severity, and the scale has a theoretical raw score between 30 and 120 (54).

The subscales of Depression (BYI-D) and Anger (BYI-A) in the Beck Youth Inventories of Emotional and Social Impairment (72) each consist of 20 brief self-statement-questions on a 4- point scale ranging from 'Never' to 'Always,' each with a theoretical raw score between 0 and 60 (72). Higher scores indicate higher symptom severity. The BYI-D scale is extensively used in Swedish Child and Adolescent Depression clinics (24) and recommended by the Swedish Agency for Health Technology Assessment and Assessment of Social Services (73) to use when screening for MDD among adolescents. Cronbach's alpha in the current sample was 0.93 (95% CI [0.93, 0.45]) for BYI-D and 0.93 (95% CI [0.92, 0.94]) for BYI-A.

The World Health Organization Wellness Index (WHO-5) consists of 5 salutogenic self-statements. For instance, the first question 1 is worded, "Over the past two weeks I have felt cheerful and in good spirits." The scale uses a 6-point scale option ranging from 'All of the time' to 'At no time.' The theoretical raw score is between 0 and 25. Higher scores indicate well-being, and a total score below 13 indicates poor well-being. The scale has adequate validity both as a screening tool for depression and as an outcome measure in clinical trials and has been applied successfully across a wide range of study fields (74). It has also been validated in depressed adolescents in Sweden (75). In the current sample, Cronbach's alpha of this index was 0.87 (95% CI [0.86, 0.8]).

The Patient-Reported Outcome Measurements Information System (PROMIS) consists of item banks for various health, and lifestyle dimensions developed to advance the science and application of patient-reported outcomes. Items are worded in the past tense, for example, starting "In the last 7 days, ..." and continuing with a statement. For example, PROMIS Anxiety consists of 15 statements (e.g. "... I felt like something awful might happen"), and PROMIS Friend consists of 10 statements (e.g. "... I was able to count on my friends"). The respondent answers on a 5-point scale, with possible answers ranging from 'never' to 'almost always' (76). Theoretical score ranges are between 0 and 60 for PROMIS Anxiety and 0 and 40 for PROMIS Friend; higher scores indicate higher anxiety levels and better peer relationships. Cronbach's alpha in the current sample was 0.92 (95% CI [0.91, 0.93]) for PROMIS Anxiety and 0.92 (95% CI [0.91, 0.93]) for PROMIS Friend.

4.3.2 Data analyses

Data were first analyzed with descriptive statistics, and the Mann-Whitney U test was used to test mean differences between groups. We performed a confirmatory factor analysis (CFA) with a four-factor correlated model to test the model proposed by Reynolds and previously confirmed in other versions of the scale (77). We used chi-square (χ^2), the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error approximation (RMSEA) to test the goodness-of-fit of the model. Good fit for a given model was determined by RMSEA less than 0.06, and at least less than 0.08, TLI and CFI > 0.95 for an excellent fit and > 0.90 for acceptable fit (78-80). Since RADS-2 has ordinal scale variables with only 4-point scale options, the robust diagonal weighted least square (DWLSSS) estimator was used (81, 82).

When evaluating a self-report questionnaire, it is also essential to know whether different groups perceive the measure the same way. Otherwise, it would be possible for mean differences to occur only because the scale is perceived differently by those different groups. Using measurement invariance (MI) as a part of the factor analytic framework can test this assumption. In order to test MI, the first step is to establish a configural model for the different groups and then constrain the factor loadings, thresholds, and residuals step-by-step to find evidence for metric, scalar, and strict MI. The model fit for the metric model is compared to the model fit for the configural model, the model fit for the scalar

model is compared to the model fit for the metric model, and so on. Since the observed means equal the intercept/thresholds of the variables added to the factor loadings multiplied by the factor score, it is, in theory, possible for the intercept/thresholds to be unequal, which would result in elevated or attenuated observed means for different groups. That, in turn, would give a biased observed mean (83). When MI holds, it allows for the interpretation that the observed mean differences among groups are due to actual differences.

We tested for MI according to Svetina and Rutkowski's (84) method to establish that the scale measures are the same for both the older and younger age groups and sexes. Since chi-square is sensitive to sample size, ΔCFI and ΔRSMEA were evaluated, and also the Satorra-Bentler test was performed. If scalar measurement invariance holds, it is possible to evaluate latent mean differences. In order to establish measurement invariance, the following goodness of fit indices was used: $\Delta\text{RSMEA} \leq 0.05$ and significant $\Delta\chi 2$ and CFI $\geq$ -0.004 for metric invariance, $\Delta RSMEA \leq 0.01$ and significant $\Delta\chi 2$ and CFI $\geq$ -0.002 (84-86), as well as a negative Satorra-Bentler test indicating that the null hypothesis fails and that there is no difference between the models.

Reliability was tested with Cronbach's alpha (see Introduction for discussion of this measurement). The comparative analyses of different reliability measures showed the risk of underestimating Cronbach's alpha when assumptions are violated (60). Cronbach's alpha $\geq$ 0.7 was classified as 'acceptable,' $\geq$ 0.8 as 'good,' and $\geq$ 0.9 as 'excellent' (87).

Concurrent validity is shown if a measure of the same construct, in this case, depression, has a high correlation with the scales currently validated, and correlation estimates between RADS-2 and BYI-D tested this type of validity. Convergent validity is considered to be established when various measures of similar constructs are correlated, and this type of validity was tested with correlation estimates between RADS-2 and PROMIS anxiety, the BYI-A subscale, and WHO-5. Finally, discriminant validity tests whether a measure of a different construct discriminates to the validated scale. In this study, discriminant validity was considered to be established if the correlation between RADS-2 and PROMIS friend was moderate. Concurrent, discriminant, and convergent validity were examined with Pearson's r. Values of 0.1–0.29 were considered to be a slight

correlation, 0.3–0.49 medium correlation, and 0.50 and above were strong correlations (88).

4.4 Study III

4.4.1 Steps used to translate the PROMIS item banks into Swedish

In order to translate the eight PROMIS pediatric item banks into Swedish, authorization was obtained from PROMIS Health Organization (PHO). The translation process followed the Functional Assessment of Chronic Illness Therapy (FACIT) translation methodology (46) with some modifications (specifically the use of a review group). All item banks used unipolar verbal response scales with five response alternatives: never, almost never, sometimes, often, almost always. The PHO offers PROMIS pediatric item definition lists (PROMIS organization, 2018) aimed to help with the translation of the items. The following item banks had item definition lists: anger, anxiety, depressive symptoms, fatigue, pain interference, and peer relationships.

Semantic/linguistic, content, and conceptual adaptation was performed according to the method proposed by Vet et al. (89). Two independent forward translations were completed, and a third researcher and the original translators reconciled any differences between the two translations. This translation was then submitted to a multi-professional bi-lingual review group. The next step was to submit the questions to review groups, which is a modification of the FACIT method. The review groups provided an opportunity to get in-depth assessments from various professionals at the same time. The review group consisted of 21 persons with professions like questionnaire design experts; researchers experienced using patient-reported measures in healthcare; linguists; and pediatric healthcare professionals. The group review was conducted in a two-day session, after which back-translation was carried out. The bilingual translation team then did a final review. A PROMIS organization member (JC) reviewed each back-translated item to assess the equivalence of the source and target translation. A final report was written documenting the development of each translation. After that, cognitive debriefing interviews were carried out for all eight item banks. Four researchers were trained in conducting cognitive debriefing interviews, and a Swedish language cognitive interview manual was written and used for guidance. The interviews were performed with eleven

children aged between 8 and 17 years (9 girls and 2 boys, mean and median age 14 years). All children were fluent in Swedish. Think-aloud methodology and subsequent respondent debriefing were used (90). As a final step, the penultimate version of the questionnaire was completed by a small sample of the target population. The respondents answered the full battery of questions and were then asked how they perceived the questions and the web survey version used for the school sample in Studies II and IV.

4.4.2 Data analyses

Both inductive and deductive methodologies (content analysis(91)) were used to analyze the data qualitatively and to simultaneously quantify the data. The number of items that belonged to each subtheme was counted. It was noted when in the translation process the potential issues were discovered and how many times.

4.5 Study IV

4.5.1 Measures

PROMIS Pediatric Bank v2.0 – Anxiety consists of 15 questions, and the PROMIS Pediatric Bank v2.0 – Depressive Symptoms consists of 14 questions. They are both based on a 5-point response option scale ranging from 'never' to 'almost always' and use a seven-day recall period (76). The inferred severity of the condition of the respondents is given in theta, and PROMIS transforms thetas to T-scores through the formula (theta * 10) + 50 = T-score. Higher values indicate higher levels of the underlying construct (i.e., higher levels of anxiety or depressive symptoms).

4.5.2 Data analyses

We examined unidimensionality with Kaiser-Meyer-Olkin factor adequacy (KMO), parallel analysis, exploratory factor analysis (EFA), and single-factor confirmatory factor analysis (CFA). In order to avoid performing the EFA and CFA on the same sample (92), the sample was randomly split in half. The parallel analysis and EFA were conducted based on the polychoric correlations matrix using the weighted least square (WLS) estimation method. The KMO, parallel analysis, and EFA was performed using the R package Psych (93). The CFA was performed based on the polychoric correlations matrix with the robust diagonal least square (DWLSSS) estimation method using the R package lavaan (94). A KMO value greater than 0.9 is characterized as 'marvelous,' greater than 0.7 as 'middling,' and less than 0.5 as unacceptable (95). In the EFA, unidimensionality was assumed when the first factor accounted for at least 20% of the variability and the ratio of the variance explained by the first to the second factor was greater than four (94). In the single-factor CFA, the following fit indices were used: CFI > 0.95, TLI > 0.95, RMSEA < 0.06 and SRMR < 0.08 (78).

Secondly, we evaluated local independence. Local independence is verified when there is no significant association between item responses after the dominant factor has been controlled for. We evaluated local independence with Yen's Q3 statistics Samejima's graded response model (GRM) for polytomous items using the mirt package in R (96). Yen's Q3 < 0.02 was used to flag for LD (97).

Thirdly, monotonicity was assessed. Monotonicity is the probability of a respondent (correctly) selecting a higher response category in tandem with having a higher level of the underlying trait. Monotonicity was tested with nonparametric item response theory using the Mokken package in R (98). Monotonicity was evaluated with the scalability coefficient H > 0.30 for items and > 0.50 for the item banks (98).

After IRT model assumptions were evaluated and items that lacked local independence were deleted, the item banks were again fitted within the item response theory framework using Samejima's graded response model (GRM) for polytomous items using the mirt package in R (96). GRM yields the slopes and the threshold values of the items. The item slopes refer to the discriminative

ability of the item, where a higher value indicates a better discriminative ability. The item thresholds refer to the item difficulty, and for a 5-point option scale, four thresholds are located along with the measured trait. Fits for the items were evaluated with Orlando and Thissen's S-X^2 statistics, where a non-significant value is an indication of adequate fit ($p > 0.001$) (99).

In order to ensure equivalent measurement between groups, measurement invariance was evaluated with differential item functioning (DIF). There are two types of DIF, namely uniform and non-uniform. Uniform DIF is when the magnitude of the difference between the groups is the same throughout the whole continuum of the trait, and non-uniform DIF is when the magnitude of the difference between the groups varies at various levels of the trait. DIF was evaluated for sex (girls vs. boys) and age groups (12–15 years vs. 16–20 years) with ordinal logistic regression using the Lordif package in R (100) using a McFadden's pseudo-R^2 change of 2% as a critical value to flag for DIF.

Reliability was evaluated with information that is inversely related to the standard error of the estimated construct or theta level. Information (I) and subsequently the standard error (SE) differed across the continuum of theta. Theta was estimated based on the GRM model and ranged from approximately -4 to 4. A SE of 0.548 corresponded to the reliability of 0.70, and a SE of 0.316 corresponded to the reliability of 0.90.

4.5.2.1 Computer adaptive testing (CAT)

The standard PROMIS stopping rule is a standard error of 0.3, and for pediatric PROMIS item banks, 0.4. Both standard error values were tested to find the best balance between a low standard error of measurement and the number of items administered. The starting item was the item with the highest information value for the average level of participants in the population (theta=0) according to PROMIS practice. The catR package in R was used for the simulations (95). The stopping rule of a maximum of 12 items was not used because it is not possible to condition on both standard error and the maximum number of items in catR. The maximum Fisher information criterion was used for item selection, and to estimate expected thetas, *a posteriori* estimation (EAP) was used.

In order to avoid performing GRM and CAT on the same sample, we randomly split the sample into two. The first sample (the 'evaluation sample') was used for the GRM and gave the calibration item parameters to use in the CAT. The second sample (the 'validation sample') was used for the response matrix in the CAT simulation.

4.6 Ethical considerations

For study I, all data had been collected before the start of the thesis. The Research Ethics Committees at Uppsala University and the Regional Ethical Review Board in Umeå approved the study. The questionnaire questions may have been perceived as intrusive, but the risk and possible negative consequences for the participants in this project were considered low. The researchers had no access to personal data.

The Regional Ethical Review Board at Umeå University approved studies II-IV and studies II and IV, was also approved by the principals of each school. All participants received written and oral information about the study and gave written consent prior to participating. School participants < 15 years old also provided written parental consent. All participants were given a unique code that was only saved in a USB stick, separated from their names, and stored in a locked archive only accessible by the Principal Investigator. The participants were reimbursed after the first and second tests with a gift ticket. No adverse events were expected except that perhaps completing the assessments might have been perceived as tiring, tedious, or frustrating and that some of the questions may have made the adolescents feel uncomfortable. Participants were free to decline to answer any questions, leave the questionnaire incomplete, and ask the research assistant for clarifications if needed. We think that the benefits of conducting the study outweigh any potential risks for the participants.

5. Summary of studies

5.1 Study I

The primary aim of study I was to investigate whether there have been changes in self-reported mental health symptoms among adolescents. We compared two geographically identical groups of 16-year-olds at two time-points, 1981 and 2014, with regard to self-reported internalized and externalized mental health symptoms. In 1981, the sample comprised 1083 respondents (506 girls and 577 boys), and in 2014, the sample comprised 682 respondents (338 girls and 344 boys). The response rate was 99.7% in 1981 and 98.3% in 2014. In 1981, the number of parents born outside the Nordic countries and the number of respondents with divorced parents was significantly lower than in 2014, whereas the number of parents with blue-collar jobs was significantly higher in 1981 than in 2014 (Table 1).

Table 1. Demographic data describing the two samples, 1981 and 2014, with sample size (n), percentage (%), and z-value for comparison between the groups.

Variable	**1981 n (%)**	**2014 n (%)**	**z-value**
Parental origin			-9.51**
Both Nordics	989 (98.3)	594 (87.0)	
One parent Nordic	16 (1.6)	44 (6.4)	
Neither Nordic	1 (0.1)	45 (6.6)	
Parental occupational classification			-8.84**
Blue collar	494 (49.1)	206 (30.0)	
White and blue collar	331 (32.9)	250 (36.4)	
White collar	181 (18.0)	231 (33.6)	
Living arrangements			-5.01**
With mother and father	782 (78.0)	459 (67.0)	
Single parent or other	221 (22.0)	226 (33.0)	
Parental unemployment			-0.91 (p=0.37)
Unemployed	193 (19.3)	142 (21.1)	
Employed	809 (80.7)	532 (78.9)	

Note. Mann–Whitney U test for comparison between groups. Z-values were significant at $p < 0.001$. **$p < 0.001$.

Regarding internalized symptoms, we found that anxiety, depressive, and functional somatic symptoms (FSS) were significantly higher in 2014 than in 1981 for both girls and boys (Table 2), and at both time points, girls scored significantly higher on questions asking about symptoms than boys. The interaction effect for sex was significant, indicating that the slope of the increase was different in girls than in boys. In 2014, the difference between girls and boys was greater than in 1981 (Figure 6). Regarding externalized symptoms, we found that conduct problems were significantly lower in 2014 than in 1981 for both girls and boys, and the differences seen between girls and boys in 1981 were no longer seen in 2014.

Table 2. Mean and standard deviations for depressive and anxiety symptoms, functional somatic symptoms, and conduct problems sorted by gender and year. Between-group differences (girls/boys and 1981/2014) are shown separately for each symptom category.

	1981			**2014**			**Boys**	**Girls**
	Total M (SD)	Boys M (SD)	Girls M (SD)	Total M (SD)	Boys M (SD)	Girls M (SD)	Total M (SD)	Total M (SD)
Depressive symptoms	0.48 (0.34)	0.40 (0.34)	0.55 (0.31)	0.68 (0.41)	0.53 (0.37)	0.82 (0.40)	0.45 (0.36)	0.66 (0.38)
n	1007	526	481	682	344	338	870	819
p, df =1, 1685	<0.001[a] F=46.23			<0.001[a] F=113.65			<0.001[b] F=28.77	<0.001[b] F=114.48
Anxiety symptoms	0.12 (0.23)	0.07 (0.16)	0.17 (0.29)	0.32 (0.45)	0.17 (0.29)	0.48 (0.52)	0.11 (0.22)	0.30 (0.43)
n	1006	526	480	680	343	337	869	817
p, df=1, 1682	<0.001[a] F=22.02			<0.001[a] F=161.73			<0.001[b] F=18.48	<0.001[b] F=189.19
FSS	0.33 (0.25)	0.29 (0.25)	0.37 (0.25)	0.57 (0.35)	0.46 (0.31)	0.67 (0.36)	0.36 (0.28)	0.50 (0.33)
n	1007	526	481	682	344	338	870	819
p, df=1, 1685	<0.001[a] F=19.62			<0.001[a] F=96.93			<0.001[b] F=69.77	<0.001[b] F=220.44
Conduct problems	2.03 (1.47)	2.51 (1.46)	1.52 (1.31)	1.21 (1.28)	1.26 (1.36)	1.15 (1.19)	2.02 (1.54)	1.37 (1.27)
n	992	514	478	659	330	329	844	807
p, df=1, 1647	<0.001[a] F=134.10			0.286[a] F=1.14			<0.001[b] F=172.81	<0.001[b] F=14.69

Note. Depressive and anxiety symptoms and FSS 0–2p. Conduct problems 0–5p. Between values calculated with two-way ANOVA. [a] denotes significant differences between gender. [b] denotes significant differences between year.

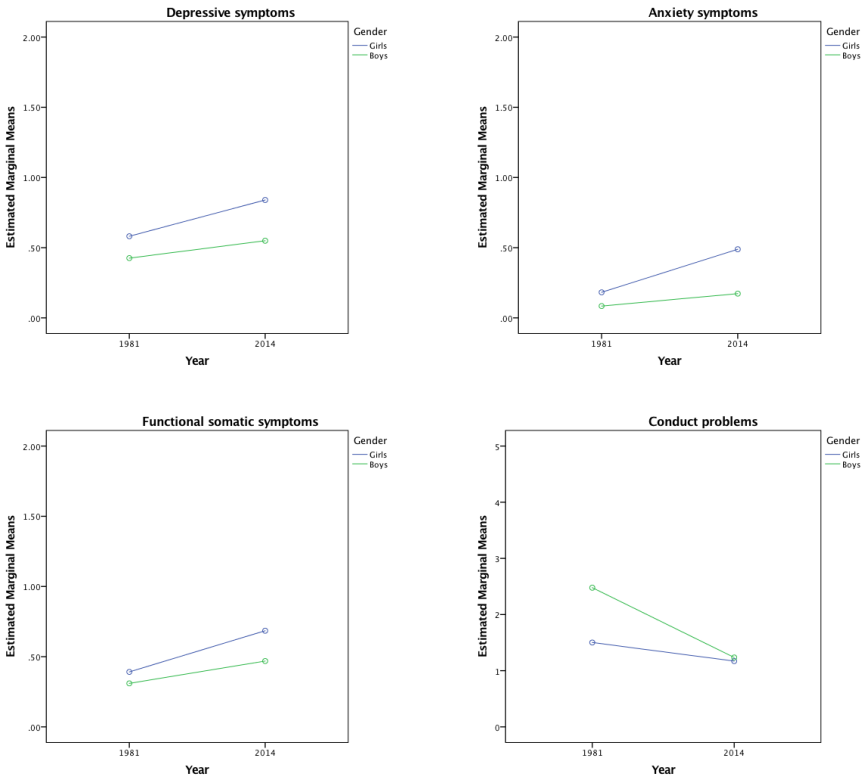

Figure 6. Estimated marginal means adjusted for parents' occupational classifications, parents' country of origin, parents' employment status, and adolescent's living arrangements for 1981 and 2014 in girls and boys for depressive and anxiety symptoms, functional somatic symptoms (FSS), and conduct problems. The interaction effect between gender and year was significant in all cases ($p < 0.001$).

5.2 Study II

This study aimed to test the psychometric properties of the Swedish version of the RADS-2 in a normative sample. RADS-2 is an internationally established multi-dimensional measure of adolescent depression that is compatible with the DSM and ICD systems and has not yet been translated and used in Sweden.

Reliability was tested with Cronbach's alpha, which ranged from acceptable (0.77, anhedonia/negative affect subscale) to excellent (0.93, RADS-2 total scale) for the subscales and total scale. Correlations between RADS-2 and BYI-D ranged from 0.56 (anhedonia/negative affect) to 0.88 (RADS-2 total scale), thus indicating concurrent validity. Convergent validity was shown by correlation to PROMIS – Anxiety ranging from 0.43 (anhedonia/negative affect) to 0.70 (RADS-2 total scale), by correlation to BYI-A ranging from 0.46 (anhedonia/negative affect) to 0.73 (RADS-2 total scale), and by correlation to WHO-5 ranging from 0.48 (anhedonia/negative affect) to 0.72 (RADS-2 total scale). Discriminant validity was shown by correlation to PROMIS Friend ranging from -0.38 (somatic complaints) to -0.50 (anhedonia/negative affect). These findings are in line with previous findings that self-assessment measures of depression are strongly associated with scores on related internalizing and psychosocial measures such as anxiety and low self-esteem (101, 102). The model fit for the correlated 4-factor model for the total sample (sensitive to sample size) was significant (χ^2 (399) = 1738.61, $p < 0.001$), but the other fit indices were found to be acceptable (CFI = 0.945, TLI = 0.940, RMSEA = 0.072 (90% CI [0.069 – 0.076])). Thus, the confirmatory factor analysis supported the 4-factor structure proposed by Reynolds (54) with acceptable fit indices. Measurement invariance was confirmed using the guidelines of Svetina and Rutkowski (84). RADS-2 was found to be invariant across both sex (girls, boys) and age groups (12–15 years, 16–20 years), making it a useful measure for interpreting gender and age differences in the assessment of depression symptoms. Table 3 presents this.

Table 3. Measurement Invariance Goodness of Fit for the 4-Factor Model of Reynolds Adolescent Depression Scale second edition (RADS-2) presented by sex and age-group.

Invariance (sex)	χ^2(df)	CFI	TLI	RSMEA (90% CI)	$\Delta\chi$	Δdf	ΔCFI	ΔRSMEA
Configural	1865.697** (798)	0.951	0.947	0.065 [0.061 – 0.069]				
Metric	1907.590** (828)	0.950	0.948	0.064 [0.060 – 0.068]	41.893	30	-0.001	-0.001
Scalar	1908.320** (854)	0.952	0.951	0.062 [0.059 – 0.066]	0.730	26	0.002	-0.002
Invariance (age group)	χ2(df)	CFI	TLI	RSMEA (90% CI)	$\Delta\chi$	Δdf	ΔCFI	ΔRSMEA
Configural	1984.697** (798)	0.948	0.944	0.068 [0.065 – 0.072]				
Metric	2023.640** (828)	0.948	0.945	0.067 [0.064 – 0.071]	38.943	30	0	-0.001
Scalar	2012.682** (854)	0.949	0.949	0.065 [0.061 – 0.069]	-10.958	26	0.001	0.002

Note. ** p < 0.001. χ2 = Chi square; df = degrees of freedom; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; TLI = Tucker-Lewis Index; ΔCFI = Change in Comparative Fix Index. For sex groups (girls and boys), Satorra-Bentler $\Delta\chi^2$: Config vs. Metric χ^2 (30) = 38.425, p = 0.139. Metric vs. Scaler χ^2 (26) = 35.428, p = 0.103. For age groups: Satorra-Bentler $\Delta\chi^2$: Config vs. Metric χ^2 (30) = 26.83, p = 0.6322. Metric vs. Scaler χ^2 (26) = 22.736, p = 0.648.

5.3 Study III

This study aimed to translate and culturally adapt eight pediatric PROMIS item banks into Swedish. The item banks were anger, anxiety, depressive symptoms, family relationships, fatigue, pain interference, peer relationships, and physical activity, and contained 116, out of which 24 items presented translational problems to be resolved. We categorized the translation problems into three themes: 1. Lack of matching definitions with items across languages (6 items), 2. Problems related to language, vocabulary, and cultural differences (6 items), and 3. Difficulties in adaption to age-appropriate language (12 items). Each theme was then categorized into subthemes. See Table 4 for the themes and subthemes with examples.

Table 4. Themes of translation issues.

Main themes	Subthemes	Example of issues (number of items)
Lack of matching definitions with items across languages	Equivocal items with precise definitions	Anxiety item bank: 'I was afraid of going to school', where the item could mean both being in school and traveling to school. The item was translated in accordance to the item definition* and the concept of being in school. (5)
	Equivocal items without precise definitions	Physical activity item bank: 'How many days did you run for 10 minutes or more?'. Whether it refers to 10 minutes of continuous running or 10 minutes in total during the day is not clear. The item was translated to 10 minutes of continuous running. (1)
Problems related to language, vocabulary and cultural differences	Adjectival agreement on intensity levels of the concept to be translated	Anxiety item bank: The translation of the word 'scared' used in the Anxiety item bank was translated to 'rädd'. The word 'rädd', can be back-translated to either 'scared' or 'afraid'. In Swedish, the grade difference between afraid and scared is more difficult to clearly illustrate using only a single word. (2)
	Culturally specific idiomatic phrases	Physical activity item bank: 'How many days did you exercise or play so hard that your muscles burned?'. 'Muscles burned' is an idiomatic phrase in English and this could not be translated directly to Swedish. Instead, this was translated to 'How many days did you exercise or play so much that you got aching muscles?'. (3)
	Cultural differences of measurements	Pain interference item bank: 'It was hard for me to walk one block when I had pain'. The informal American English measurement of 'one block' has no Swedish equivalent. The translation must therefore relate to either the exact distance of a block (if such an exact measure exists) or an approximation. The Swedish translation used both: 'it was difficult for me to walk a short distance (about 100 meters) when I was in pain'. (1)

Difficulties in adaption to age-appropriate language	Comprehensibility of the items changes in the translation process	Anger item bank: I felt upset' the most precise translation of the item 'I felt upset' in the Anger item bank was 'upprörd' which was difficult to understand for children of younger age. Hence, this item was instead translated to 'I felt both angry and sad'. (4)
	Acceptance of the items for all age groups	Peer relationships item bank: 'Other kids wanted to be with me' was not accepted by teenagers until 'kids' was replaced by 'others my age'. This phrase was then reused in other items for consistency and thus avoided the problem of having to use the word 'child'. (8)

Note. *Definition list from PROMIS organization (2018).

5.4 Study IV

Study IV aimed to describe the IRT analysis of the Pediatric PROMIS item banks of anxiety and depressive symptoms in a sample of Swedish adolescents and a Child and Adolescent Psychiatry (CAP) sample. A total of 637 students (mean age 15.73 (SD = 1.76) 61.1 % girls) and 291 patients (mean age 15.64 (SD =1.61) 71.4 % girls) participated in the study.

We found that both item banks showed sufficient unidimensionality and monotonicity. In order to achieve local independence, two anxiety items and three depressive symptoms items were removed. Subsequently, a graded response model was fit to the item banks. Discriminative values gave high precision for all items, and threshold values ranged from -0.47 to 3.28 for the anxiety item banks and -1.24 to 2.00 for the depressive symptoms item banks. No DIF was found for sex, age group, or sample type. Mean thetas for the sample were 50.0 (SD = 9.5) for the anxiety item bank and 50.01 (SD = 9.6) for the depressive symptom item bank. Figure 7 presents the density plot for the theta distribution for the depressive symptoms and anxiety item banks.

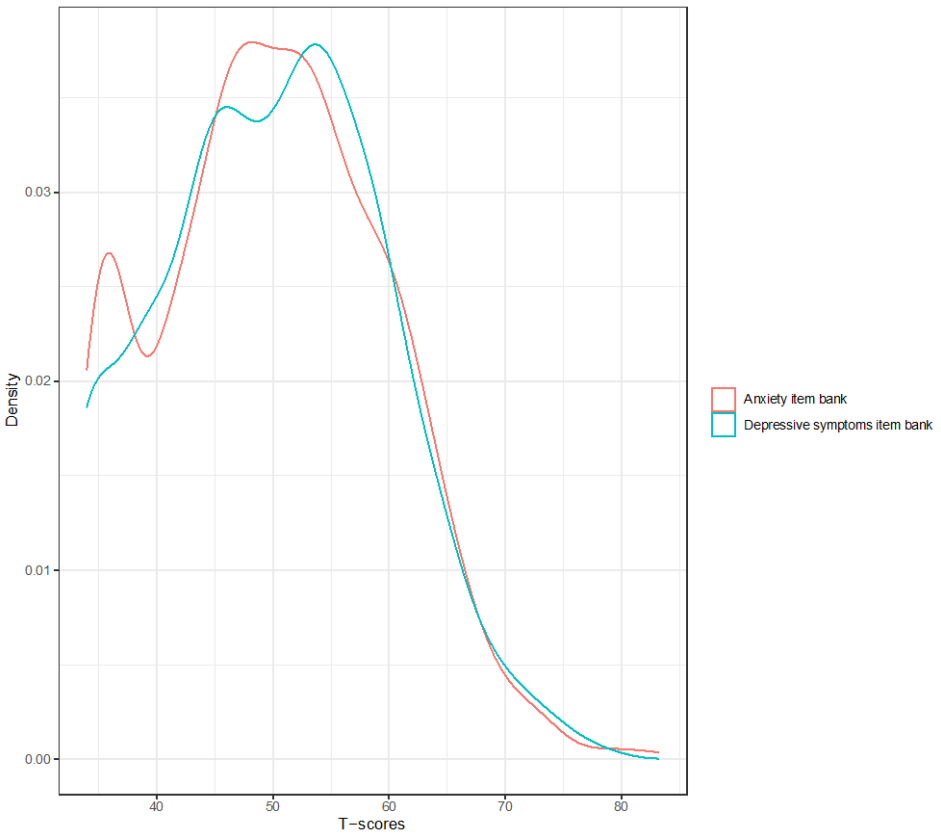

Figure 7. T-score distribution for the Swedish pediatric Patient-Reported Outcomes Measurement Information System (PROMIS) anxiety and depressive symptoms item banks

A CAT simulation showed high reliability (> 0.90) for both item banks between T-scores of approximately 45 to 75. At the lowest end of T-scores, the standard error of measurement was above the threshold of 0.3 (at a T-score of less than 45, the reliability is poor). However, these values are within the "normal" limits of symptoms, which means they are less likely to be useful in a CAP setting. At higher T-scores, the standard error of measurement and number of items are lower, indicating higher precision at higher levels of anxiety and depressive symptoms (Figures 8 and 9).

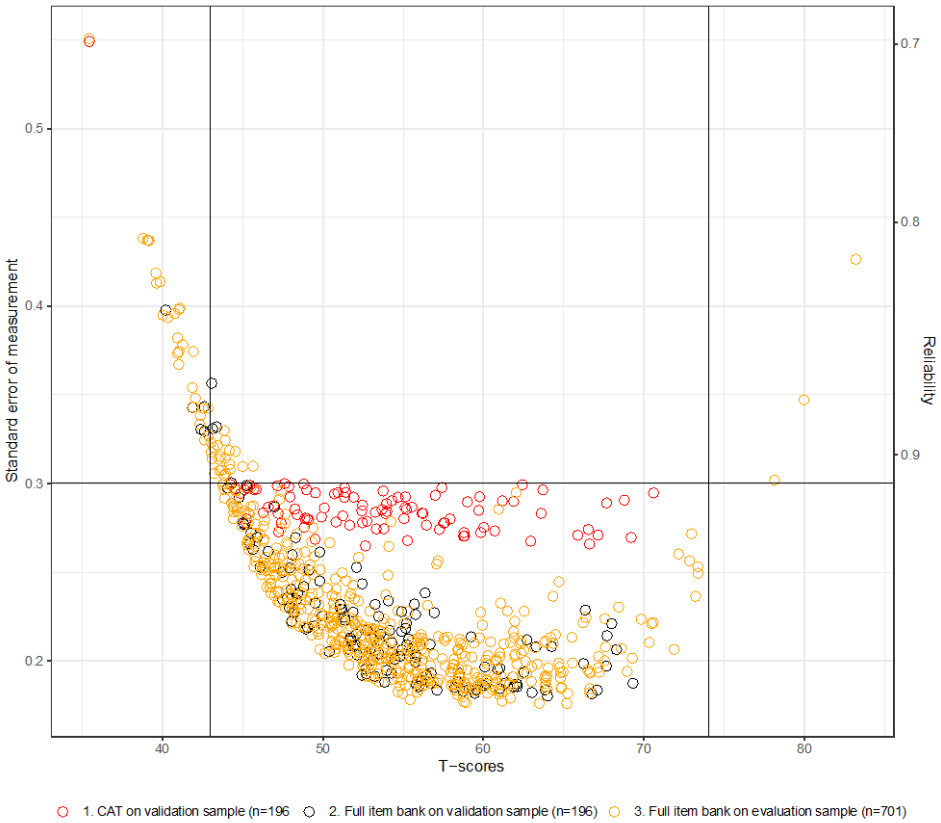

Figure 8 Reliability of the Swedish pediatric PROMIS anxiety item bank comparing CAT with full item banks on the validation and evaluation samples.

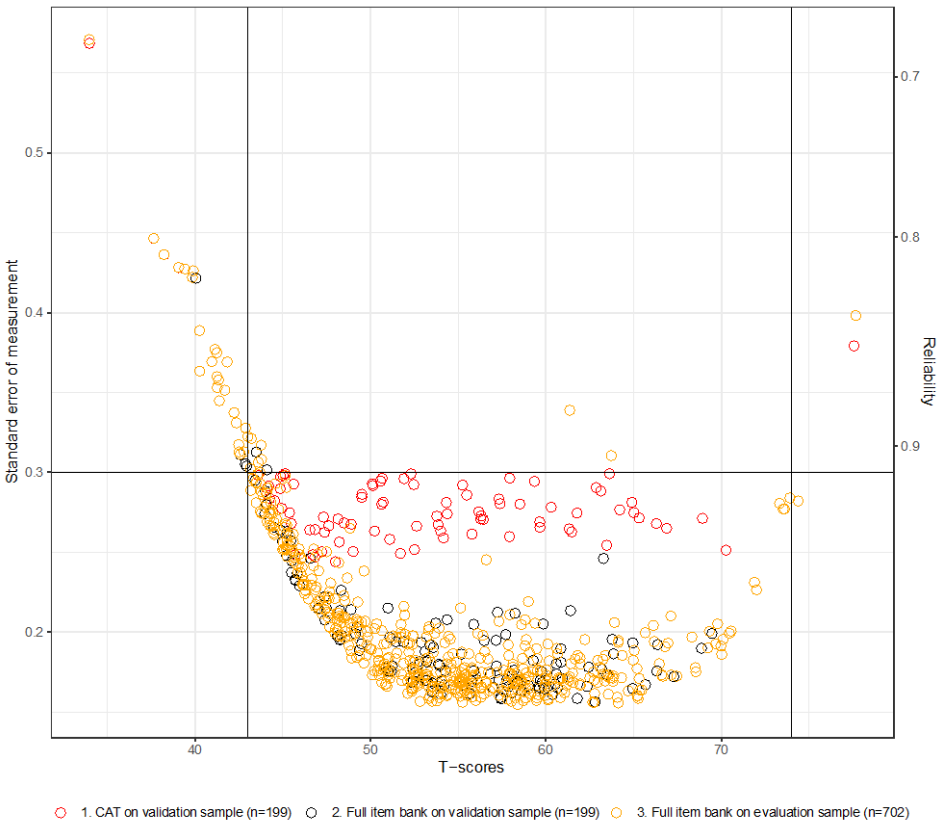

Figure 9 Reliability of the Swedish pediatric PROMIS depressive symptoms item bank comparing CAT with full item banks on the validation and evaluation samples

6. Discussion

6.1 Summary of time trends in the mental health of adolescents

In recent decades, self-reported mental health symptoms in adolescents seem to have increased, and reviews suggest a minor increasing trend in this age group (1, 3, 103). During the last two years, the Covid-19 pandemic and the measures taken to limit the spread of the virus seem to have helped accelerate this increase (104, 105). In recent years in Sweden, it has also been a dramatic increase in the number of children and adolescents seeking CAP care (106).

Our study of grade 9 students in Luleå (which took place before the Covid-19 pandemic) showed the same trend as noted in other studies, with increasing internalized symptoms, especially among girls. Below, we propose a way to contextualize the observed time trends regarding youth mental health in order to theorize about the reasons behind the increase of internalized mental health symptoms.

6.2 Possible societal explanations for the rise of mental health symptoms in the young

6.2.1 Child development according to Urie Bronfenbrenner

Urie Bronfenbrenner (1917-2005), a developmental psychologist, developed the ecological systems theory of child development (107). In this model, he put the individual child/youth in the center and described various levels of systems around that individual. The level closest to the child/youth was called the microsystem. The microsystem contains the immediate relations of the child, the people that the child has direct interactions with, such as family, friends, teachers, and classmates. The second level is the mesosystem, which includes the interactions between agents in the child's microsystem, for example, between divorced parents or between the child's teachers and parents. The third level is the exosystem, which includes factors outside the child's intermediate contact,

such as the parents' working environments or the neighborhood. The fourth level, the macrosystem, is the overall cultural, political and societal system in which the child lives, for example, Sweden, a country situated in northern Europe with democracy.

Last but not least is the chronosystem, which is the overall impact of events that may affect a child's trajectory, such as a parent's death, a pandemic, geopolitical changes, or climate change. All systems have direct or indirect interactions with each other, and therefore affect the development of the child/youth at the center; as Bronfenbrenner stated regarding ecogical research, "…the principal main effects are likely to be interactions." (107, p. 38).

6.2.2 Societal changes in Sweden

It may be perceived as paradoxical that mental illness in adolescents increased in Sweden since Sweden is a country where poverty and unemployment are not as prevalent as in many other countries. Sweden is a relatively well-organized, safe, and democratic welfare state and a society with functioning and accessible schools and healthcare systems. However, have there been changes in the micro, meso-, exo-, macro-and chronosystems during recent years that could explain the observed trends. If we start from the periphery, with a chronosystem analysis focusing on larger societal patterns, we can see changes such as climate change, the unraveling of ecosystems and increased urbanization, and (since March 2020) the Covid-19 pandemic. Several of these changes are driven by a global financial system built on unlimited economic growth on a finite planet and externalized costs of production (108). Lately, it seems that the mental health implications of the pending climate crisis (109) have dawned on us (110).

These changes have coincided with an exponential technological development that has changed our society immensely. Social media networks such as Instagram, Snapchat, and Facebook are new phenomena that reach many adolescents worldwide, perhaps as frequently as daily. The emerging studies in this field have shown associations between the use of these modalities and the increase of internalizing symptoms (111, 112).

The globalized economy and a more neoliberal value system have also brought on changes at the macrosystem level. During the last 30 years in Sweden, there has been a systemic shift from a social-democratic welfare state to a partly regulated neoliberal-based society in which individual capacity for self-regulation, discipline, and control is highly valued (113). Simultaneously, employees have less control and greater pressure to perform (113). These factors have been proposed to lead to increased individual stress and vulnerability, and they possibly correspond to a marked increase of stress-related disorders, which is now considered a significant public health concern, especially among women (114). Additionally, the "me-too movement" has uncovered aspects of a persistent gender order, in which men's domination is created and maintained, hidden structural and institutionalized sexual violence against women, together affecting women's mental health negatively (115-117).

On the exosystem level, over the last thirty years, we have seen changes in the labor market, such as increased job insecurity, more temporary employment, and increased work-related stress (118). On the meso- and microsystem levels, the school system has undergone alterations such as decentralization and educational marketization (119, 120). At the same time, Sweden has experienced a decline in school results (121). There are reports of differences in reading ability depending on socioeconomic factors in grade 4 students in Sweden (122). Differences between the educational quality and content concerning socio-demographic, gender, and ethnicity have also been established during this period (123).

It is possible to place several additional factors known to be adding risk for mental health symptoms in both adolescents and adults in the macro-, exo-, meso- and microsystems, for instance, poor sleep (124, 125), dietary habits (126, 127) and obesity (128). Studies indicate time trends worsening sleep habits with a decrease in sleep duration (129, 130), possibly due to the availability and use of media devices such as smartphones at night (131). There are tendencies in Europe and the US with more obesity among adolescents (128, 132). Studies show the correlations between sedentary behavior and depressive symptoms (133) and the interplay between sedentary behavior and unhealthy diet (134). A recent study also shows correlations between outdoor artificial light and mental health among adolescents (135) and correlations with screen time and depressive symptoms (136). There are also inequalities in the housing markets and evidence that areas

with lower socioeconomic status are more exposed to noise and air pollution (137, 138), affecting mental and physical health (139-143). There are also differences depending on socioeconomic factors on eating habits (144, 145).

The levels in the ecological systems that Bronfenbrenner described interact with each other. Therefore, changes in the outer circles will ripple into the epicenter, with a cumulative effect on children and youth. Bronfenbrenner's theory challenges the belief that mental health problems in the young are purely medical disorders to be treated with medication or psychological problems treated with psychotherapy.

A way of adding even more layers to Bronfenbrenner's model is to incorporate theories of intersectionality. In order to make visible how factors such as gender, sexual orientation, age, race, and socioeconomic status, may have additional adverse effects on mental health (146-148).

6.3 The relevance of self-report measures in Child- and Adolescent Psychiatry

There are yet no objective measures upon which to base psychiatric diagnoses. On a physiological level, magnetic resonance imaging, heart rate variability, and an array of blood biomarkers have been studied in conjunction with these measures. However, the relationships are not clear cut, and none of these methodologies are used as diagnostic tools in clinical settings (149-152). Therefore, the use of questionnaires is an even more important feature than for many other physical diseases that have objective biomarkers in addition to patient symptoms. For example, rheumatoid arthritis (RA) is a disease with objective and subjective markers, such as a blood test showing high inflammatory parameters and a positive RA-factor besides patient symptoms such as fatigue, joint pain, and stiffness (153). In psychiatry, such clear connections between biomarkers and symptoms have yet to be found.

Therefore, self-reported measures are used to draw conclusions on several levels, such as patient, patient-group, and population levels. At the patient level, scales are used as diagnostic support and follow treatment outcomes (73, 154-156). Self-reported measures are also used on a patient-group level, such as the evaluation of clinical trials of medications or psychotherapy (157, 158). Most clinical trials are randomized controlled trials (RCTs). RCTs are considered the highest standard of intervention research, and therefore RCT results will have a high impact on clinical guidelines and treatment recommendations. Two recent studies from the Netherlands have shown the advantages of using IRT fit indices instead of sum scales when evaluating RCT results (159, 160). Using IRT scales lead to less bias (159, 160). Self-reported measures are also used to assess mental health development and time trends, which often constitute the basis of health policies and the distribution of resources (161).

6.3.1 Reflections on scale evaluation

A psychometric evaluation of a scale can potentially be a subjective task, such as in the case of confirmatory factor analysis (CFA), where there are many different estimators (162, 163). The maximum-likelihood estimator was the first to be used for CFA, and the equation for the maximum-likelihood estimator rests upon the assumption that the variables are continuous. When CFA started to be used in psychological research with ordinal data, other estimators were "developed," such as the robust maximum-likelihood and diagonally weighted least squares (81, 82, 164). Most psychological measures have ordinal variables and, therefore, should use ordinal estimators in the CFA. However, ordinal estimators might lead to inflated fit indices, which erroneously result in an acceptable model fit (165).

In CTT, there are also the disadvantages noted in the background above. For example, the problem of the generalizability beyond the sample, technically the parameter estimates should not be generalized beyond the sample (56, 57); the use of Cronbach's as the primary reliability measure (60); the lack of knowledge about the possibility of evaluating systematic errors; the assumption about normally distributed responses at an item level, often not fulfilled with psychological measures which can result in biased parameter estimates (166). Fortunately, the PROMIS initiative seems to have worked at least as a partial bridge in this respect since there is a proposed protocol (31) to follow for evaluating the scale with IRT, which improves the overall quality. Some of the

risks with deficient self-report measures can be counteracted with more precise scales. The use of IRT techniques makes it possible to perform differential item functioning (DIF) (100), which enables comparison of it the items are perceived differently depending on group prerequisites such as sex or age group. If the scale is invariant between groups hence has no DIF, then differences can be interpreted as such (100). However, scales validated with IRT also have some limitations, such as the difficulty of administering IRT-based tests, the lack of the necessary expertise to perform such analysis, and the interpretation of factor scores. Furthermore, CAT needs to be done on a computer, and both IRT and CAT have sample sizes recruitments, demanding more extensive samples than CTT for validation.

Finally, even if a scale is reliable and valid, what do we know of the diagnostic system used as the template for questionnaires.

6.4 Diagnostic systems

The Diagnostic and Statistical Manual of Mental Disorders (DSM) system is a diagnostic system classifying clusters of symptoms into different diagnoses. The first version was printed in 1952, and since then, there have been five versions. Currently, most CAP clinics in Sweden base their diagnostics on DSM-V. The system was intended to make psychiatric diagnoses comparable and more similar to other medical disciplines by applying a structured system., The system is well established and frequently used, but critique against DSM is also prevalent (167-170). As mentioned above, one problem with DSM-based psychiatric diagnoses is that they are only descriptions of symptoms that are clustered together into a diagnostic entity, and there is no empirical way of knowing if any particular clustering of symptoms is more meaningful in terms of reflecting underlying pathologies than any other (171). The National Institute of Health in the US (NIH) has recently started to move away from the DSM system in favor of the Research Domain Criteria (RDoC) framework.

RDoC is an initiative of the National Institute of Mental Health (NIMH), and this framework was created to aid the development of individualized precision medicine for mental health and “help identify new targets for treatment

development, detect subgroups for treatment selection, and provide a better match between research findings and clinical decision making" (172). The structure of the RDoC framework is described as "a matrix in which the rows represent various constructs grouped hierarchically into broad domains of function" in which "the columns of the matrix denote different levels of analysis, from genetic, molecular, and cellular levels, proceeding to the circuit-level…", "… and on to the level of the individual, family environment, and social context" (172). However, the RDoC system still assumes that psychiatric and mental health symptoms have neurobiological and biobehavioral correlations. It has been pointed out that the assumption that mental suffering is a medical disorder continues to favor a biological treatment approach, a view which might risk not acknowledging the impact of the social context (168).

As a reaction to the problems mentioned above with the diagnostic classification systems, the power threat meaning (PTM) framework has been developed by the British Psychological Society as a meta-framework that uses several different models, practices, and philosophical traditions for an alternative diagnostic approach. The aim is "to inform and expand existing approaches by offering a fundamentally different perspective on the origins, experience, and expression of emotional distress and troubled or troubling behavior" (168). The framework acknowledges that humans are social beings with core needs such as experiencing a sense of justice and fairness within their community, having a sense of security and belonging to a family and social group, being safe, valued, accepted, and loved in their earliest relationships with caregivers. Unfulfilled core needs can be experienced as a threat to physical, emotional, relational health and safety, and survival. The framework further describes how the dynamics of, e.g., biological, legal, economic, and cultural capital power ultimately affect all humans. These power dynamics can interfere on an individual or a community level and give rise to varying individual or collective narratives and formulations of meaning. The combination of power, threat, and meaning-making can lead to threat responses currently defined within DSM systems as 'psychiatric symptoms' leading to psychiatric diagnoses, when they are perhaps better understood as stress responses, that is, as acute or chronic reactions to threat or as adaptive and rational survival strategies. The weight and importance of intersectionality in understanding and treating mental health problems are heavily emphasized in the PTM-framework.

6.5 Limitations

During the work with this thesis, it became apparent that there are several limitations within the current research paradigm of psychiatry. As noted by Borsboom et al., the underlying measure needs to exist for a measure to have construct validity (in the term, it is incorporated the assumption that the measure captures an actual underlying disease). “If something does not exist, then one cannot measure it” (173). Mental health symptoms and the suffering they entail surely do exist but are the current diagnostic systems most commonly used the most accurate way of describing them. If there is a validity problem incorporated in the diagnostic system, that will also have implications on the psychometrics evaluations of scales (174). Psychometric evaluation and methodology will only be as strong as its weakest link.

The following limitations from the individual studies should also be noted: Study 1: A limitation may be that behavioral problems may be expressed differently in 2014 compared to the 1980s and therefore not captured the same way at the two time points. Study 2: A limitation of this study is that we did not gather data on ethnicity only on nationality in general and therefore could not validate the RADS-2 in ethnic minority groups. Furthermore, we relied on self-report measures alone as validation measures for RADS-2. Participants were not geographically stratified and did not match the Swedish general pediatric population. Instead, the participants constituted a convenience sample drawn from four different schools from different socioeconomic areas. Study 3: A possible limitation was that the cognitive debriefing took place in a single geographical area; hence, there is a risk of not reflecting the dialectal variations of Sweden. Richer data might have been obtained by including a more extensive representation of boys, younger children, and children from more diverse linguistic groups in the cognitive debriefing. Study 4: A limitation of this study was that we lacked appropriate ways to evaluate whether the respondents answered truthfully. They could have answered entirely at random without consideration of the item as such or their actual feelings. Such respondents can be hard to detect. One possible way to detect them is by adding reverse-coded items and deleting those with conflicting answers(175). Also, because of the GDPR jurisdiction, the online platform had to be changed during the data collection. We, therefore, collected some of the data in paper format. However, research on other scales has not shown significant differences whether the data has been adhered to online or in paper format (176).

6.6 Strengths

In addition to the limitation named above, these studies also have several strengths. In study I, the geographic setting was stable, an identical questionnaire was used at both time points, the scope of the total population in the defined age range was extensive, and there was an extraordinarily high response rate. In Study II, strengths included the evaluation of generalizability of scores on this measure over time (i.e., test-retest reliability) and collecting data from different schools from different socioeconomic areas. In Study II, expert review groups emanating from already-established pediatric healthcare networks throughout the country were used in the translation process. In Study IV, strengths included the inclusion of both school and patient samples and IRT use.

6.7 Summary and future directions

The objective of this thesis was to add to our knowledge regarding time trends of self-reported mental health among Swedish youth and to validate internationally-used and reliable self-report measures for use in Sweden. Study I was a cross-sectional study on two different samples of grade 9 students in Luleå county, 33 years apart, answering the same self-report measures. Results from this study confirmed previous research regarding a rise of internalizing mental health symptoms, especially among girls (it should be noted that the second time point was before the Covid-19 pandemic). In Study II, an internationally-established measure for depression (RADS-2), not previously validated in Sweden, was validated in a Swedish school sample with classical test theory. To further contribute to the field, the exploration continued from classical to modern test theory. Study III described the translation and cultural adaption of eight pediatric PROMIS item banks into Swedish (the PROMIS started as an initiative to advance patient-reported outcomes and has clear benefits such as the use of item response theory). Study IV used item response theory to validate two pediatric PROMIS item banks, anxiety and depressive symptoms, in a Swedish school and patient sample.

On a personal note, during the work with this thesis, the importance of being cautious in the interpretation of scales became apparent. A scale can always be interpreted qualitatively as a way of learning more about the respondent.

However, quantitative interpretation demands the knowledge that the scale is generalizable to the patient or patient group in question. Caution is also needed due to the inadequacies of the diagnostic system (168). As such, the implementation of PROMIS item banks in Swedish CAP clinics would advance the overall quality of measures. Last but not least, the work on this thesis has made me conscious of the importance of considering intersectional factors and societal changes when viewing the increase of mental health problems among youth.

From a historical perspective, over the last century, there has been an explosion in knowledge in the area of mental health, but there are still major knowledge gaps in our contemporary paradigms. As discussed by the philosopher Jonna Bornemark, this current explosion in knowledge may have led to a less-developed relationship to complexity and less acceptance of the unknown. That, in turn, may lead to an oversimplification of reality and premature conclusions of cause and effect (177). In her book "*Det omätbaras renässans*" ("The renaissance of the immeasurable"), Bornemark discusses our time through the eyes of Renaissance philosophers. In the process, it becomes clear that we live in a cultural and social context that puts a major emphasis on and faith in measurements, classification, and categorization (177). Consequently, we have a less-developed relationship to the narrative of the subjective experience, the qualitative, and the unknown. The human desire for ordering and categorizing, the contemporary belief in these categories, and dissociation from the relationship with all that is unknown can lead to loss of perspective. A strictly medical diagnostic approach will undoubtedly miss out on aspects of a human's life that cannot be fully explained, categorized, or measured.

In summary, a future direction in child and adolescent psychiatry research from a psychometric perspective is that using item response theory and the concept of underlying ability may contribute to a better understanding of mental health symptoms. A shift in the current diagnostic approach and view on mental health, such as proposed by the Power Threat Meaning Framework, would be welcome, as would incorporation of contextual factors and their effect on mental health and theories such as intersectionality and Bronfenbrenner's ecological systems theory.

7. Acknowledgments

First, I want to start by acknowledging my grandparents Gun and Stig Blomqvist, who unfortunately has passed away, for the advice I received on my gratulation card after med-school "Fortsätt vara nyfiken!" – "Keep on being curious!" Curiosity has led me through these four years. The feeling of constantly learning and the favor of being a Ph.D. student where it's written down on your Web-ISP that you're supposed to take these super interesting courses - on your work time. I can't write this without also stating that it has been a struggle sometimes. But, to stand on the door to greater knowledge is a true indulgence.

This thesis would not have been possible without the involvement of a lot of people. Foremost, thank you to all the respondents in the studies. Without your participation, there would be no studies. I also want to send my warmest thank you to the following:

My main supervisor Eva Henje. I learned so much during these four years, and I'm thankful you took me on as a Ph-D student. Thanks for the endless corrections and read-throughs on my work and interesting discussion about research, mental health, politics, etc.

My co-supervisor Inga Dennhag, for being a safe harbor in times of need. Your structure, way of organizing your work, and drive have been inspirational and taught me a lot.

Bruno Hägglöf, who sadly has passed away. He was the supervisor on my CAP residency research project, and through him, I got in contact with Anne Hammarström.

My co-supervisor Anne Hammarström for giving me the possibility to analyze, write and be part of your research group during my work with the Luleå cohort. Your work with the Northern Swedish cohort is truly incredible.

Aiden Loe, psychometrician at Cambridge. It has been great working with you. You are such a skilled yet humble expert. My co-authors, besides my supervisors, Evalill Nilsson, John Chaplin, and Erik Ekbäck, I enjoyed working with you. To Linda Haldner Henriksson for being involved in both the half-time and kappaseminars. And Marie Wiberg for reading my kappa and IRT seminars. Thanks also to my examinator Jussi Jokkinen.

The former and present Clinical Directors on CAP Västernorrland, Lena Berglund Friberg and Mats Gidlund, the Clinical Director on CAP Örnsköldsvik, Patrik

Järnberg, and my former residency supervisor Mia Törnqvist; for believing in me and giving me the opportunity to do this work.

All my colleagues, both former and present, on BUP Örnsköldsvik. I've learned so much from working with you all. Your engagement, hard work, joyous spirits, and striving to do the best for the patients are the greatest inspiration.

FUI Västernorrland, especially Jeanette Sundberg, for making doctoral studies easier, e.g., providing financial support, forums to meet, and train tickets.

Past and present colleges in the Clinical Science and CAP department at Umeå University for help with a little bit of everything (Maud Normark and Birgitta Bäcklund) as well as interesting discussion on journal clubs. To Frida Rindestig, for friendship, interesting discussions about doctoral studies, research, CAP etc.

Olov Rolandsson for encouraging me to take the research course "Grundläggande forskningsutbildning…" which got me into research. And Marjalisa Byhamre for encouraging me to start the "Grundläggande forskningsutbildning…" course when I felt unsure and that I would be out of place. I'm glad we took it together.

All my friends old and new: my tjej-gäng that I have had the pleasure of knowing since forever, my dance buddies through Kulturskolan, study friends and family friends. And especially Lisa, who's been my closest "college" during these years, my go-to person for when in need and high stress for dwelling work things, kids, and life in general.

To my dear friend Johanna Pålsson for being you!

To my brother Joakim and my sister-in-law Johanna and their kids, you are such essential constants in our life. To the Thelin family, especially my parents-in-law Kalle and Eva, for their help and support.

To my parents for their love, help, and support always. My father grew up at a bakery. After school, he would have a pile of oven plates to hand wash. From that, he learned to take them one plate at a time. There have been *a lot of plates* during my doctoral studies, and thanks to you, mum, and dad, I have taken them one at a time.

What I'm most grateful for in life is my family. My husband Tomas and our three children, Juni, Stig, and Maj. Without the support from you, Tomas, I would never have been able to juggle the world of parenthood, residency, and doctoral studies. So, to Tomas, Juni, Stig, and Maj, you are the best! I love you!

8. References

1. Bor W, Dean AJ, Najman J, Hayatbakhsh R. Are child and adolescent mental health problems increasing in the 21st century? A systematic review. Aust N Z J Psychiatry. 2014;48(7):606-16.

2. Collishaw S. Annual research review: Secular trends in child and adolescent mental health. J Child Psychol Psychiatry. 2015;56(3):370-93.

3. Potrebny T, Wiium N, Lundegard MM. Temporal trends in adolescents' self-reported psychosomatic health complaints from 1980-2016: A systematic review and meta-analysis. PLoS One. 2017;12(11):e0188374.

4. Collishaw S, Maughan B, Goodman R, Pickles A. Time trends in adolescent mental health. J Child Psychol Psychiatry. 2004;45(8):1350-62.

5. Collishaw S, Maughan B, Natarajan L, Pickles A. Trends in adolescent emotional problems in England: a comparison of two national cohorts twenty years apart. J Child Psychol Psychiatry. 2010;51(8):885-94.

6. Hill J, Maughan B. Conduct disorders in childhood and adolescence. Cambridge, U.K.; New York, NY: Cambridge University Press; 2001. xiv, 581 p. p.

7. Wangby M, Magnusson D, Stattin H. Time trends in the adjustment of Swedish teenage girls: a 26-year comparison of 15-year-olds. Scand J Psychol. 2005;46(2):145-56.

8. West P, Sweeting H. Fifteen, female and stressed: changing patterns of psychological distress over time. J Child Psychol Psychiatry. 2003;44(3):399-411.

9. Carlson J. Health Behaviour in School-aged Children (HBSC), results from Sweden of the 2013/14 WHO study. In: Sweden TPHAo, editor.: Grafisk form: AB Typoform; 2014.

10. Petersen S, Bergström E, Cederblad M, Ivarsson A, Köhler L, Rydell A, et al. Barns och ungdomars psykiska hälsa i Sverige. En systematisk litteraturöversikt med tonvikt på förändringar över tid Stockholm (eng: The mental health of children and adolescents in Sweden. A systematic literature

review with an emphasis on changes over time). Kungl Vetenskapsakademien, Hälsoutskottet. 2010.

11. The Royal Swedish Academy of Sciences KV. Trender i barns och ungdomars psykiska hälsa i Sverige (Trends in children's and young people's mental health in Sweden). State-of-the-Science Konferensuttalande. 2010.

12. Eaton WW, Neufeld K, Chen LS, Cai G. A comparison of self-report and clinical diagnostic interviews for depression: diagnostic interview schedule and schedules for clinical assessment in neuropsychiatry in the Baltimore epidemiologic catchment area follow-up. Arch Gen Psychiatry. 2000;57(3):217-22.

13. Goodman R. The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. J Child Psychol Psychiatry. 1999;40(5):791-9.

14. Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJ, et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. PLoS Med. 2013;10(11):e1001547.

15. Mathers C, Fat DM, Boerma JT. The global burden of disease : 2004 update. Geneva, Switzerland: World Health Organization; 2008.

16. Thapar A, Collishaw S, Pine DS, Thapar AK. Depression in adolescence. Lancet. 2012;379(9820):1056-67.

17. Birmaher B, Ryan ND, Williamson DE, Brent DA, Kaufman J. Childhood and adolescent depression: a review of the past 10 years. Part II. J Am Acad Child Adolesc Psychiatry. 1996;35(12):1575-83.

18. Korczak DJ, Goldstein BI. Childhood onset major depressive disorder: course of illness and psychiatric comorbidity in a community sample. The Journal of pediatrics. 2009;155(1):118-23.

19. Socialstyrelsen. Psykisk ohälsa bland unga (Mentall illness among young people). Underlagsrapport till Barns och ungas hälsa, vård och omsorg 2013. The National Board of Health and Welfare; 2013. 2013-5-43.

20. Jensen PS, Rubio-Stipec M, Canino G, Bird HR, Dulcan MK, Schwab-Stone ME, et al. Parent and Child Contributions to Diagnosis of Mental

Disorder: Are Both Informants Always Necessary? Journal of the American Academy of Child & Adolescent Psychiatry. 1999;38(12):1569-79.

21. Smith SR. Making Sense of Multiple Informants in Child and Adolescent Psychopathology: A Guide for Clinicians. Journal of Psychoeducational Assessment. 2007;25(2):139-49.

22. Socialstyrelsen. Upptäcka psykisk ohälsa hos barn och ungdomar – En sammanställning av systematiska översikter (Detecting mental illness in children and adolescents); 2013. ISBN 978-91-7555-016-9. 2013-18.

23. Socialstyrelsen. Barn- och ungdomspsykiatrins metoder. En nationell inventering. (Methods of child and adolescent psychiatry. A national inventory). Västerås; 2009. 2009-126-146.

24. Dunerfeldt, M. Elmund, A. Söderström, B. Bedömningsinstrument inom BUP i Stockholm. Kartläggning och faktasammanställning. Barn- och ungdomspsykiatri. Stockholms läns landsting.; 2010. ISSN 1653-204X.

25. NIH. NIoH. NIH Launches Second Phase of Patient Reported Outcomes Initiative 2009 [Available from: https://www.nih.gov/news-events/news-releases/nih-launches-second-phase-patient-reported-outcomes-initiative.

26. NIH. NIoHOoSC-TCF. Patient-Reported Outcomes Measurement Information System (PROMIS) - Program Snapshot [Website]. [updated January 29, 2019. Available from: https://commonfund.nih.gov/promis/index.

27. Alonso J, Bartlett SJ, Rose M, Aaronson NK, Chaplin JE, Efficace F, et al. The case for an international patient-reported outcomes measurement information system (PROMIS®) initiative. Health and quality of life outcomes. 2013;11(1):210-.

28. Cella D, Gershon R, Lai J-S, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation - Official Journal of the International Society of Quality of Life Research. 2007;16(Supplement 1):133-41.

29. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS)

developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol. 2010;63(11):1179-94.

30. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther. 2011;13(5):R147.

31. Reeve BB, Hays DR, Bjorner BJ, Cook FK, Crane KP, Teresi AJ, et al. Psychometric Evaluation and Calibration of Health-Related Quality of Life Item Banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care. 2007;45(5 Suppl 1):S22-S31.

32. Fries J, Rose M, Krishnan E. The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. J Rheumatol. 2011;38(8):1759-64.

33. Magasi S, Ryan G, Revicki D, Lenderking W, Hays RD, Brod M, et al. Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. Qual Life Res. 2012;21(5):739-46.

34. Fussner L, Black W, Lynch-Jordan A, Morgan E, Ting TV, Kashikar-Zuck S. Utility of the PROMIS Pediatric Pain Interference Scale in Juvenile Fibromyalgia. Journal Of Pediatric Psychology. 2019;44(4):436-41.

35. Askew RL, Cook KF, Revicki DA, Cella D, Amtmann D. Evidence from diverse clinical populations supported clinical validity of PROMIS pain interference and pain behavior. Journal of Clinical Epidemiology. 2016;73:103-11.

36. DeWalt DA, Gross HE, Gipson DS, Selewski DT, DeWitt EM, Dampier CD, et al. PROMIS® pediatric self-report scales distinguish subgroups of children within and across six common pediatric chronic health conditions. Quality of life research. 2015;24(9):2195-208.

37. Hinds PS, Nuss SL, Ruccione KS, Withycombe JS, Jacobs S, DeLuca H, et al. PROMIS pediatric measures in pediatric oncology: Valid and clinically feasible indicators of patient-reported outcomes. Pediatric Blood & Cancer. 2013;41(6):402-8.

38. Yeatts KB, Stucky B, Thissen D, Irwin D, Varni JW, Dewitt EM, et al. Construction of the Pediatric Asthma Impact Scale (PAIS) for the Patient-Reported Outcomes Measurement Information System (PROMIS). Journal of Asthma. 2010;47(3):295-302.

39. Devine J, Klasen F, Moon J, Herdman M, Hurtado MP, Castillo G, et al. Translation and cross-cultural adaptation of eight pediatric PROMIS® item banks into Spanish and German. Quality of life research. 2018;27(9):2415-30.

40. Haverman L, Grootenhuis MA, Raat H, van Rossum MA, van Dulmen-den Broeder E, Hoppenbrouwers K, et al. Dutch-Flemish translation of nine pediatric item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS)(R). Qual Life Res. 2016;25(3):761-5.

41. Liu Y, Wang J, Hinds PS, Wang J, Shen N, Zhao X, et al. The emotional distress of children with cancer in China: an item response analysis of C-Ped-PROMIS Anxiety and Depression short forms. Quality of life research. 2015;24(6):1491-501.

42. Vilagut G, Forero CG, Castro-Rodriguez J, Abellanas A, Alonso J. Calibration and validation of the Spanish version of PROMIS depression using a population sample. Qual Life Res. 2016;25(s1):104-.

43. Pinto MNFdC, Pinto RdMC, Mendonça TMdS, Souza CG, da Silva CHM. Validation and calibration of the patient-reported outcomes measurement information system: Pediatric PROMIS® Emotional Distress domain item banks, Portuguese version (Brazil/Portugal). Quality of life research. 2020;29(7):1987-97.

44. de Castro NFC, de Melo Costa Pinto R, Da Silva Mendonça TM, Da Silva CHM. Psychometric validation of PROMIS® Anxiety and Depression Item Banks for the Brazilian population. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2020;29(1):201.

45. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. Value in health. 2005;8(2):94-104.

46. Eremenco SL, Cella D, Arnold BJ. A Comprehensive Method for the Translation and Cross-Cultural Validation of Health Status Questionnaires. Evaluation & the health professions. 2005;28(2):212-32.

47. Acquadro CMD, Conway KMA, Hareendran AP, Aaronson NP. Literature Review of Methods to Translate Health-Related Quality of Life

Questionnaires for Use in Multinational Clinical Trials. Value in health. 2008;11(3):509-21.

48. Raykov T, Marcoulides GA. Introduction to psychometric theory. New York: Routledge; 2011.

49. Martinez-Martin P. Composite rating scales. Journal of the neurological sciences. 2009;289(1):7-11.

50. Cappelleri JCP, Jason Lundy JP, Hays RDP. Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures. Clinical therapeutics. 2014;36(5):648-62.

51. Furr M, Bacharach VR. Psychometrics: An Introduction2007.

52. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin. 1955;52(4):281-302.

53. Drost EA. Validity and Reliability in Social Science Research. Education Research and Perspectives. 2011;38(1):105-23.

54. Reynolds WM. Reynolds Adolescent Depression Scale: Professional manual (2nd ed). Odessa, FL: Psychological Assessment Resources, Inc.; 2002.

55. Petrillo J, Cano SJ, McLeod LD, Coon CD. Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to Evaluate Patient-Reported Outcome Measures: A Comparison of Worked Examples. Value in health. 2015;18(1):25-34.

56. Theresa K. Classical Test Theory: Assumptions, Equations, Limitations, and Item Analyses. Thousand Oaks: SAGE Publications, Inc; 2005. p. 91.

57. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, N.J. ; London: L. Erlbaum Associates; 2000.

58. Streiner DL, Norman GR. Health Measurement Scales : A practical guide to their development and use. 4th ed. Oxford: Oxford University Press; 2008.

59. DeVellis RF. Classical Test Theory. Medical care. 2006;44(11):S50-S9.

60. McNeish D. Thanks coefficient alpha, we'll take it from here. Psychological methods. 2018;23(3):412.

61. Raykov T, Marcoulides GA. Thanks Coefficient Alpha, We Still Need You! Educational and Psychological Measurement. 2019;79(1):200-10.

62. Liu B, Engstrom K, Jadback I, Ullman S, Berman AH. Child self-report and parent ratings for the Strengths and Difficulties Questionnaire: Norms and agreement in a Swedish random population sample.(Research Article). Scandinavian Journal of Child and Adolescent Psychiatry and Psychology. 2017;5(1):13.

63. Zumbo BD, Gadermann AM, Zeisser C. Ordinal versions of coefficients alpha and theta for likert rating scales. Journal of Modern Applied Statistical Methods. 2007;6(1):21-9.

64. Chalmers RP. On Misconceptions and the Limited Usefulness of Ordinal Alpha. Educational and Psychological Measurement. 2018;78(6):1056-71.

65. Hancock GR, & Mueller, R. O. Rethinking construct reliability within latent variable systems. In R. In: Cudeck SdT, & D. Sörbom (Eds). editor. Structural equation modeling: Present and future - A festschrift in honor of Karl Jöreskog. Lincolnwood, IL.: Scientific Software International.; 2001.

66. Choi SW, Swartz RJ. Comparison of CAT Item Selection Criteria for Polytomous Items. Applied psychological measurement. 2009;33(6):419-40.

67. Weiss DJ. Improving Measurement Quality and Efficiency with Adaptive Testing. Applied psychological measurement. 1982;6(4):473-92.

68. Fries JF, Bruce B, Cella D. The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. Clinical and Experimental Rheumatology. 2005;23(5):S53-S7.

69. Statistikmyndigheten SCB. Swedish socio-economic classification (SEI). https://www.scb.se/contentassets/22544e89c6f34ce7ac2e6fefbda407ef/english_ov9999_1982a01_br_x110p8204-3.pdf

; 1984.

70. Statistikmyndigheten SCB. SEI, yrkesförteckning version 2020-04-16. https://www.scb.se/contentassets/22544e89c6f34ce7ac2e6fefbda407ef/sei_index_webb_20200416.pdf; 2020.

71. Hammarström A, Westerlund H, Kirves K, Nygren K, Virtanen P, Hägglöf B. Addressing challenges of validity and internal consistency of mental health measures in a 27- year longitudinal cohort study - the Northern Swedish Cohort study. BMC Medical Research Methodology. 2016;16:4.

72. Beck JS, Beck, A. T., Jolly, J. Manual for the Beck Youth Inventories of Emotional and Social Impairment: San Antonio, TX: The Psychological Corporation.; 2001.

73. SBU. Diagnostik och uppföljning av förstämningssyndrom. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2012. SBU-rapport nr 212. ISBN 978-91-85413-52-2. In: (SBU) Sbfmu, editor. 2012.

74. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: A Systematic Review of the Literature. Psychotherapy and psychosomatics. 2015;84(3):167-76.

75. Blom EH, Bech P, Hogberg G, Larsson JO, Serlachius E. Screening for depressed mood in an adolescent psychiatric context by brief self-assessment scales--testing psychometric validity of WHO-5 and BDI-6 indices by latent trait analyses. Health And Quality Of Life Outcomes. 2012;10(1):149.

76. Irwin D, Stucky B, Langer M, Thissen D, DeWitt E, Lai J-S, et al. An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation - Official Journal of the International Society of Quality of Life Research. 2010;19(4):595-607.

77. Hyun MS, Nam KA, Kang HS, Reynolds WM. Reynolds Adolescent Depression Scale - Second Edition: initial validation of the Korean version. J Adv Nurs. 2009;65(3):642-51.

78. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 1999;6(1):1-55.

79. Hooper D, Coughlan J, Mullen MR. Structural equation modelling: Guidelines for determining model fit. Electronic Journal of Business Research Methods. 2008;6(1):53-60.

80. Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. Psychol Bull. 1980;88(3):588-606.

81. Li C-H. The Performance of ML, DWLS, and ULS Estimation With Robust Corrections in Structural Equation Models With Ordinal Variables. Psychological Methods. 2016;21(3):369-87.

82. Li C-H. Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. Behav Res Methods. 2016;48(3):936-49.

83. Kite BA, Jorgensen TD, Chen P-Y. Random Permutation Testing Applied to Measurement Invariance Testing with Ordered-Categorical Indicators. Structural Equation Modeling: A Multidisciplinary Journal. 2018;25(4):573-87.

84. Svetina D, Rutkowski L, Rutkowski D. Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using Mplus and the lavaan/semTools Packages. Structural Equation Modeling. 2019;27(1):<xocs:firstpage xmlns:xocs=""/>.

85. Rutkowski L, Svetina D. Measurement Invariance in International Surveys: Categorical Indicators and Fit Measure Performance. Applied Measurement In Education. 2017;30(1):39-51.

86. Svetina D, Rutkowski L. Multidimensional Measurement Invariance in an International Context: Fit Measure Performance With Many Groups. J Cross Cult Psychol. 2017;48(7):991-1008.

87. Nunnally JC, Bernstein IH. Psychometric theory . 3. ed. New York: McGraw-Hill; 1994.

88. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Academic Press; 2013.

89. Vet H, Terwee C, Mokkink L, Knol D. Measurement in medicine: a practical guide. Vet HCWd, Terwee CB, Mokkink LB, Knol DL, editors. Cambridge: Cambridge University Press; 2011.

90. Irwin DE, Varni JW, Yeatts K, DeWalt DA. Cognitive interviewing methodology in the development of a pediatric item bank: a patient reported outcomes measurement information system (PROMIS) study.(Research)(Report). Health and Quality of Life Outcomes. 2009;7:3.

91. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. Nursing & health sciences. 2013;15(3):398-405.

92. Fokkema M, Greiff S. How Performing PCA and CFA on the Same Data Equals Trouble. European journal of psychological assessment : official organ of the European Association of Psychological Assessment. 2017;33(6):399-402.

93. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research. . Northwestern University, Evanston, Illinois R package version 219, . 2021.

94. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. J Stat Softw. 2012;48(2):1-36.

95. Kaiser HF. An index of factorial simplicity. Psychometrika. 1974;39(1):31-6.

96. Chalmers RP. mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software. 2012;48(6).

97. Chen W-H, Thissen D. Local Dependence Indexes for Item Pairs Using Item Response Theory. Journal of educational and behavioral statistics. 1997;22(3):265-89.

98. van der Ark LA. Mokken scale analysis in R. Journal of statistical software. 2007;20(11):1-19.

99. Orlando M, Thissen D. Further Investigation of the Performance of S - X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. Applied psychological measurement. 2003;27(4):289-98.

100. Choi SW, Gibbons LE, Crane PK. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. Journal of statistical software. 2011;39(8):1-30.

101. Esposito C, Clum G. Specificity of Depressive Symptoms and Suicidality in a Juvenile Delinquent Population. Journal of Psychopathology and Behavioral Assessment. 1999;21(2):171-82.

102. Smith M, Calam R, Bolton C. Psychological Factors Linked to Self-Reported Depression Symptoms in Late Adolescence. Behavioural and Cognitive Psychotherapy. 2009;37(1):73-85.

103. Li J-Y, Li J, Liang J-H, Qian S, Jia R-X, Wang Y-Q, et al. Depressive Symptoms Among Children and Adolescents in China: A Systematic Review and Meta-Analysis. Medical science monitor. 2019;25:7459-70.

104. Thorisdottir IE, Asgeirsdottir BB, Kristjansson AL, Valdimarsdottir HB, Jonsdottir Tolgyes EM, Sigfusson J, et al. Depressive symptoms, mental wellbeing, and substance use among adolescents before and during the COVID-19 pandemic in Iceland: a longitudinal, population-based study. The Lancet Psychiatry. 2021;8(8):663–72.

105. Luijten MAJ, van Muilekom MM, Teela L, Polderman TJC, Terwee CB, Zijlmans J, et al. The impact of lockdown during the COVID-19 pandemic on mental and social health of children and adolescents. Quality of life research. 2021.

106. Uppdrag Psykisk Hälsa. Psykiatrin i siffror. Barn- och ungdomspsykiatri - Kartläggning 2020. (Psychiatry in numbers. Child and adolescent psychiatry). Stockholm; 2020.

107. Bronfenbrenner U. The ecology of human development : experiments by nature and design. Cambridge, Mass.: Harvard University Press; 1979. xv, 330 p. p.

108. Steffen W, Rockström J, Richardson K, Lenton TM, Folke C, Liverman D, et al. Trajectories of the Earth System in the Anthropocene. Proceedings of the National Academy of Sciences of the United States of America. 2018;115(33):8252.

109. Masson-Delmotte V, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu and B. Zhou (eds.). IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change 2021.

110. Clemens V, von Hirschhausen E, Fegert JM. Report of the intergovernmental panel on climate change: implications for the mental health policy of children and adolescents in Europe—a scoping review. European child & adolescent psychiatry. 2020;

111. Rosenthal SR, Buka SL, Marshall BDL, Carey KB, Clark MA. Negative experiences on Facebook and depressive symptoms among young adults. Journal of Adolescent Health. 2016;59(5):510-6.

112. Maras D, Flament MF, Murray M, Buchholz A, Henderson KA, Obeid N, et al. Screen time is associated with depression and anxiety in Canadian youth. Preventive Medicine. 2015;73(C):133-8.

113. Larsson B, Letell M, Thörn Hk. Transformations of the Swedish welfare state : from social engineering to governance? New York: Palgrave Macmillan; 2012. xii, 322 p. p.

114. Försäkringskassan., Agency. SSI. Sjukfrånvaro i psykiatriska diagnoser. En registerstudie av Sveriges arbetande befolkning i åldern 20–69 år. (Sick leave in psychiatric diagnoses. A register study of Sweden's working population aged 20-69). 2020 (Socialförsäkringsrapport 2020:8).

115. Zetterström D, Landstedt, Gillander G. Depressive symptoms and the associations with individual, psychosocial, and structural determinants in Swedish adolescents. 2012;4(10):881-9.

116. Zetterström Dahlqvist H, Landstedt E, Almqvist YB, Gillander Gadin K. A non-randomised pragmatic trial of a school-based group cognitive-behavioural programme for preventing depression in girls. Int J Circumpolar Health. 2017;76(1):1396146.

117. Lange E, Young S. Gender-based violence as difficult knowledge: pedagogies for rebalancing the masculine and the feminine. International journal of lifelong education. 2019;38(3):301-26.

118. Berglund TH, K.; Isidorsson, T.; Alfonsson, J. Temporary Employment and the Future Labor Market Status. Nordic Journal of Working Life Studies. 2017;7(2).

119. Lundahl L, Arreman IE, Holm A-S, Lundström U. Educational marketization the Swedish way. Education enquiry. 2013;4(3):497-517.

120. Román H, Hallsén S, Nordin A, Ringarp J. Who governs the Swedish school? Local school policy research from a historical and transnational curriculum theory perspective. Nordic journal of studies in educational policy. 2015;2015(1):27009.

121. Gustafsson J-E, Blömeke S. Development of School Achievement in the Nordic Countries During Half a Century. Scandinavian journal of educational research. 2018;62(3):386-406.

122. Skolverket. PIRLS 2016 Läsförmågan hos svenska elever i årskurs 4 i ett internationellt perpektiv. (PIRLS 2016 The reading ability of Swedish students in grade 4 in an international perpektiv). Stockholm; 2017. 17:1560.

123. Skolverket. Vad påverkar resultaten i svensk grundskola? Kunskapsöversikt om betydelsen av olika faktorer. (What affects the results in Swedish primary schools? Knowledge overview on the importance of different factors). Stockholm; 2009. 09:1127.

124. McCallum SM, Batterham PJ, Calear AL, Sunderland M, Carragher N, Kazan D. Associations of fatigue and sleep disturbance with nine common mental disorders. Journal of Psychosomatic Research. 2019;123:109727.

125. Shochat T, Cohen-Zion M, Tzischinsky O. Functional consequences of inadequate sleep in adolescents: A systematic review. Sleep medicine reviews. 2013;18(1):75-87.

126. O'Neil A, Quirk SE, Housden S, Brennan SL, Williams LJ, Pasco JA, et al. Relationship between diet and mental health in children and adolescents: a systematic review. Am J Public Health. 2014;104(10):e31-42.

127. Quirk SE, Williams LJ, O'Neil A, Pasco JA, Jacka FN, Housden S, et al. The association between diet quality, dietary patterns and depression in adults: a systematic review. BMC Psychiatry. 2013;13:175.

128. Branca F, Nikogosian H, Lobstein T. The challenge of obesity in the WHO European region and the strategies for response : summary. Copenhagen: World Health Organization, Regional Office for Europe; 2007.

129. Matricciani L, Olds T, Petkov J. In search of lost sleep: Secular trends in the sleep time of school-aged children and adolescents. Sleep medicine reviews. 2011;16(3):203-11.

130. Pallesen S, Hetland J, Sivertsen B, Samdal O, Torsheim T, Nordhus IH. Time trends in sleep-onset difficulties among Norwegian adolescents: 1983-2005. Scandinavian journal of public health. 2008;36(8):889-95.

131. Carter B, Rees P, Hale L, Bhattacharjee D, Paradkar MS. Association Between Portable Screen-Based Media Device Access or Use and Sleep Outcomes: A Systematic Review and Meta-analysis. JAMA pediatrics. 2016;170(12):1202-8.

132. Lee HPD, Lee DPD, Guo GPD, Harris KMPD. Trends in Body Mass Index in Adolescence and Young Adulthood in the United States: 1959–2002. Journal of adolescent health. 2011;49(6):601-8.

133. Vancampfort D, Stubbs B, Firth J, Van Damme T, Koyanagi A. Sedentary behavior and depressive symptoms among 67,077 adolescents aged 12-15 years from 30 low- and middle-income countries. The international journal of behavioral nutrition and physical activity. 2018;15(1):73-9.

134. Pearson N, Biddle SJH. Sedentary behavior and dietary intake in children, adolescents, and adults: A systematic review. American journal of preventive medicine. 2011;41(2):178-88.

135. Paksarian D, Rudolph KE, Stapp EK, Dunster GP, He J, Mennitt D, et al. Association of Outdoor Artificial Light at Night With Mental Disorders and Sleep Patterns Among US Adolescents. JAMA psychiatry (Chicago, Ill). 2020;77(12):1266-75.

136. Liu M, Wu L, Yao S. Dose-response association of screen time-based sedentary behaviour in children and adolescents and depression: a meta-analysis of observational studies. British journal of sports medicine. 2016;50(20):1252-8.

137. Moss AH. Norwegian inequality in two dimensions : air pollution and income. Norwegian University of Life Sciences, Ås; 2019.

138. Fairburn J, Schüle SA, Dreger S, Hilz LK, Bolte G. Social inequalities in exposure to ambient air pollution: A systematic review in the WHO European region. International journal of environmental research and public health. 2019;16(17):3127.

139. Cerletti P, Eze IC, Schaffner E, Foraster M, Viennau D, Cajochen C, et al. The independent association of source-specific transportation noise

exposure, noise annoyance and noise sensitivity with health-related quality of life. Environment international. 2020;143:105960.

140. Dzhambov AM, Markevych I, Tilov B, Arabadzhiev Z, Stoyanov D, Gatseva P, et al. Pathways linking residential noise and air pollution to mental ill-health in young adults. Environmental research. 2018;166:458-65.

141. Eze IC, Foraster M, Schaffner E, Vienneau D, Pieren R, Imboden M, et al. Incidence of depression in relation to transportation noise exposure and noise annoyance in the SAPALDIA study. Environment international. 2020;144:106014.

142. Münzel T, Sørensen M, Schmidt F, Schmidt E, Steven S, Kröller-Schön S, et al. The adverse effects of environmental noise exposure on oxidative stress and cardiovascular risk. Antioxidants & redox signaling. 2018;28(9):873-908.

143. Yang Z, Song Q, Li J, Zhang Y, Yuan X-C, Wang W, et al. Air pollution and mental health: The moderator effect of health behaviors. Environmental research letters. 2021;16(4).

144. Moraeus L, Lindroos AK, Warensjö Lemming E, Mattisson I. Diet diversity score and healthy eating index in relation to diet quality and socio-demographic factors: Results from a cross-sectional national dietary survey of Swedish adolescents. Public health nutrition. 2020;23(10):1754-65.

145. Mattisson I. Socioekonomiska skillnader i matvanor i Sverige (Socio-economic differences in eating habits in Sweden). Livsmedelsverket. National Food Agency. 2016. 9/2016

146. Cho S, Crenshaw KW, McCall L. Toward a Field of Intersectionality Studies: Theory, Applications, and Praxis. Signs: Journal of Women in Culture and Society. 2013;38(4):785-810.

147. Bauer GR. Incorporating intersectionality theory into population health research methodology: Challenges and the potential to advance health equity. Social science & medicine (1982). 2014;110:10-7.

148. Vu M, Li J, Haardörfer R, Windle M, Berg CJ. Mental health and substance use among women and men at the intersections of identities and experiences of discrimination: Insights from the intersectionality framework. BMC public health. 2019;19(1):108-.

149. Zwolińska W, Dmitrzak-Węglarz M, Słopień A. Biomarkers in Child and Adolescent Depression. Child psychiatry and human development. 2021;

150. Goldani AAS, Downs SR, Widjaja F, Lawton B, Hendren RL. Biomarkers in autism. Frontiers in psychiatry. 2014;5:100-.

151. Koenig J, Kemp AH, Beauchaine TP, Thayer JF, Kaess M. Depression and resting state heart rate variability in children and adolescents — A systematic review and meta-analysis. Clinical Psychology Review. 2016;46:136-50.

152. Koenig J, Rash JA, Campbell TS, Thayer JF, Kaess M. A meta-analysis on sex differences in resting-state vagal activity in children and adolescents. Frontiers in physiology. 2017;8:582-.

153. Kourilovitch M, Galarza-Maldonado C, Ortiz-Prado E. Diagnosis and classification of rheumatoid arthritis. Journal of autoimmunity. 2014;48:26-30.

154. Socialstyrelsen. Barn- och ungdomspsykiatrins metoder: en nationell inventering. [Methods of child and adolescent psychiatry: a national inventory] Stockholm: Socialstyrelsen; 2009.

155. Jeffrey J, Klomhaus A, Enenbach M, Lester P, Krishna R. Self-Report Rating Scales to Guide Measurement-Based Care in Child and Adolescent Psychiatry. Child and adolescent psychiatric clinics of North America. 2020;29(4):601-29.

156. (sfbup). Sfrfrb-ou. Riktlinje depression. https://www.sfbup.se/wp-content/uploads/2017/01/SFBUPRiktlinjeDepression2014.pdf; 2014.

157. Goodyer IM, Reynolds S, Barrett B, Byford S, Dubicka B, Hill J, et al. Cognitive behavioural therapy and short-term psychoanalytical psychotherapy versus a brief psychosocial intervention in adolescents with unipolar major depressive disorder (IMPACT): a multicentre, pragmatic, observer-blind, randomised controlled superiority trial. The Lancet Psychiatry. 2017;4(2):109-19.

158. Timmerby N, Austin SF, Ussing K, Bech P, Csillag C. Family psychoeducation for major depressive disorder - study protocol for a randomized controlled trial. Trials. 2016;17(1):427-.

159. Gorter R, Fox J-P, Apeldoorn A, Twisk J. Measurement model choice influenced randomized controlled trial results. Journal of Clinical Epidemiology. 2016;79:140-9.

160. Gorter R, Fox JP, Twisk JWR. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. BMC Medical Research Methodology. 2015;15(1):55.

161. Sweden). FTPHAo. Varför har den psykiska ohälsan ökat bland barn och unga i Sverige? Utvecklingen under perioden 1985–2014. 2018.

162. Flora DB, LaBrish C, Chalmers RP. Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. Frontiers in psychology. 2012;3:55-.

163. Lei P-W, Shiverdecker LK. Performance of Estimators for Confirmatory Factor Analysis of Ordinal Variables with Missing Data. Structural equation modeling. 2020;27(4):584-601.

164. Zhao Y. The performance of model fit measures by robust weighted least squares estimators in confirmatory factor analysis. In: Lei P-W, editor.: ProQuest Dissertations Publishing; 2015.

165. Shi D, Maydeu-Olivares A. The Effect of Estimation Methods on SEM Fit Indices. Educational and psychological measurement. 2020;80(3):421-45.

166. Sheng Y, Sheng Z. Is coefficient alpha robust to non-normal data? Frontiers in psychology. 2012;3:34–34.

167. Wakefield JC. Misdiagnosing normality: Psychiatry's failure to address the problem of false positive diagnoses of mental disorder in a changing professional environment. Journal of mental health (Abingdon, England). 2010;19(4):337-51.

168. Johnstone, L. & Boyle, M. with Cromby, J., Dillon, J., Harper, D., Kinderman, P., Longden, E., Pilgrim, D. & Read, J. (2018). The Power Threat Meaning Framework: Towards the identification of patterns in emotional distress, unusual experiences and troubled or troubling behaviour, as an alternative to functional psychiatric diagnosis. Leicester: British Psychological Society.

169. Wheeler EE, Kosterina E, Cosgrove L. Diagnostic and Statistical Manual of Mental Disorders (DSM), Feminist Critiques of. Singapore: John Wiley & Sons, Ltd; 2016. p. 1-3.

170. Widiger TA, Samuel DB. Diagnostic Categories or Dimensions? A Question for the Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition. Journal of abnormal psychology (1965). 2005;114(4):494-504.

171. Hyman SE. Psychiatric Disorders: Grounded in Human Biology but Not Natural Kinds. Perspectives in biology and medicine. 2021;64(1):6-28.

172. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. The American journal of psychiatry. 2010;167(7):748.

173. Borsboom D, Mellenbergh GJ, van Heerden J. The Concept of Validity. Psychological review. 2004;111(4):1061-71.

174. Borsboom D. Psychometric perspectives on diagnostic systems. Journal of clinical psychology. 2008;64(9):1089-108.

175. Klaufus LH, Luijten MAJ, Verlinden E, van der Wal MF, Haverman L, Cuijpers P, et al. Psychometric properties of the Dutch-Flemish PROMIS ® pediatric item banks Anxiety and Depressive Symptoms in a general population. Quality of life research. 2021.

176. Hollandare F, Andersson G, Engstrom I. A comparison of psychometric properties between internet and paper versions of two depression instruments (BDI-II and MADRS-S) administered to clinic patients. J Med Internet Res. 2010;12(5):e49.

177. Bornemark J. Det omätbaras renässans : en uppgörelse med pedanternas världsherravälde. Första upplagan ed. Stockholm: Volante; 2018.