General Mental Health in Adolescence:

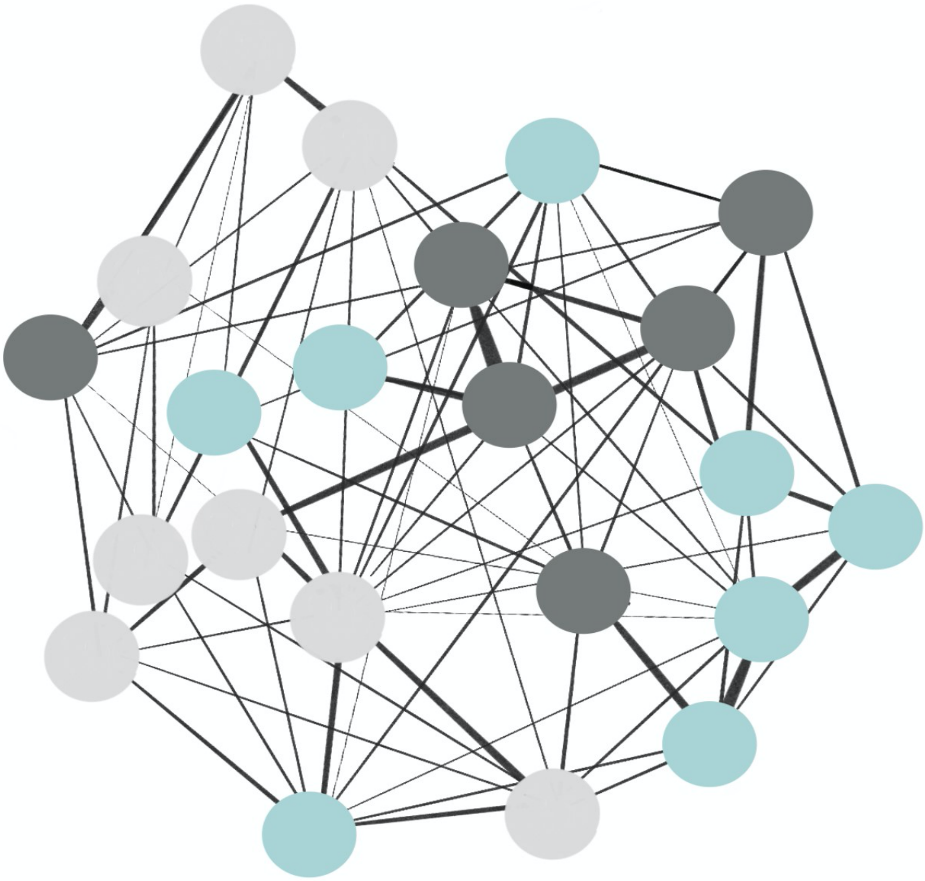
Conceptualisation and Measurement Issues

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Humanities

2022

Louise C. Black

School of Environment, Education and Development, Manchester Institute of Education



MENTAL HEALTH

IS COMPLEX.

Image by Margarita Panayiotou adapted from Paper 2 of the current thesis: https://osf.io/8jpa7/

Table of Contents

List of Tables

Because the thesis is presented in journal format, two things about this list (and that for figures) should be noted. First, tables are numbered from 1 in each of the Papers. Therefore, where necessary, the paper number is denoted in the list below by *P* in brackets. Second, where Papers are reproduced as they appear published in journals, the page of the thesis on which the paper starts is given in brackets after the journal page number.

List of Figures

Total word count for main text = 61,128

Abstract

Measuring adolescent mental health in general population samples is vital to estimating prevalence, understanding risk, and assessing intervention. Clinically rated methods suffer from ontological and reliability problems, and parent and teacher ratings typically have poor convergence with young people's self-reports. Self-reports also provide direct access to thoughts and feelings, and there are increasing calls to hear directly from young people. Despite a clear need for robust approaches to self-reported mental health in adolescence, psychometric development standards have tended to be poor.

The current thesis aimed to provide insight into more robust approaches to measuring general mental health via four papers, including secondary analysis of various datasets from the HeadStart project. Paper 1 considered the construct-level relationship between symptom and wellbeing domains. Internalizing symptoms showed equally strong relationships to wellbeing and externalizing problems, and there was also evidence for a general internalizing distress factor. Paper 2 explored indicator-level interactions over three years between internalizing symptoms, wellbeing indicators and psychosocial correlates via a multiverse framework. The multiverse framework demonstrated that the importance of indicators in the network was often sensitive to particular item operationalizations, though a few key indicators were consistently important. Paper 3 examined the age appropriateness of the commonly used self-report Strengths and Difficulties Questionnaire. Items were generally found to have inappropriate reading ages and be of low quality, while measurement invariance analysis suggested the measure functioned comparably across younger and older adolescents. Paper 4 is a meta-review of self-report general mental health measures. Content and psychometrics were analyzed. A relatively narrow range of indicators was found across constructs within mental health but measures were generally not interchangeable and had low psychometric quality.

The current thesis represents a major step forward for the theoretical and empirical understanding of general mental health. Findings suggest emotionally-focused indicators, including happiness and worry, could be particularly important when examining prevalence, risk factors, or intervention response, though more work is needed to confirm indicators with young people. The thesis also highlights a broader need to develop new measures.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgments

This thesis would not have been possible without the outstanding mentorship I have received from my two supervisors Neil Humphrey and Margarita Panayiotou. Your combined theoretical, technical, and practical expertise have been invaluable. I would also like to thank the wider MIE team, particularly Michael Wigelsworth and Pam Qualter who provided feedback and guidance at annual reviews. In addition, I would like to thank all those involved in providing, collecting, and managing data for the HeadStart project, but particularly Jess Deighton, who also provided feedback on my work. Finally, I couldn’t have done it without the support of my family and friends- thanks for all the encouragement and ad-hoc childcare!

Chapter 1: Overview of Large-Scale Measurement of Adolescent Mental Health: Problems and Opportunities

Adolescent Mental Health

Adolescence is typically considered to start with puberty and end with the transition into adult roles such as independence from parents and starting employment. As such, it is now considered to span ages 10-24 (Sawyer et al., 2018). This period is characterized by rapid physical and social change, and therefore represents distinct challenges compared to other developmental phases (Blakemore, 2019; Dahl et al., 2018; Patton et al., 2016; Sawyer et al., 2018). Issues such as puberty, changes in sleep patterns, and peer relationships are hypothesized to put young people at increased risk of mental health problems (Rapee et al., 2019). These processes are complex. While some are observable by others, the impact of, for example, changing hormones or negative relationship experiences, may only be accessible directly by the adolescent (Rapee et al., 2019). Despite these complexities, there is a clear impact in terms of mental health: Adolescents are at increased vulnerability to mental health problems, symptoms and disorders (Merikangas et al., 2010; NHS Digital, 2018; Polanczyk et al., 2015), and the majority of lifetime difficulties show first onset during this time (Jones, 2013; Solmi et al., 2021).

A substantial body of work, psychiatric epidemiology, has focused on the prevalence and etiology of mental disorders (Eaton & Merikangas, 2000), based on a medical model. Despite increasing research in this area over the last few decades (Parry-Jones, 1989; Polanczyk et al., 2015), there is some evidence that adolescent mental health is actually getting worse (Collishaw, 2015). This chapter will provide an overview of measurement in adolescent mental health, from psychiatric and other perspectives, and highlight major issues which have likely impeded scientific progress in this area. The need for more robust study of self-report data will also be demonstrated.

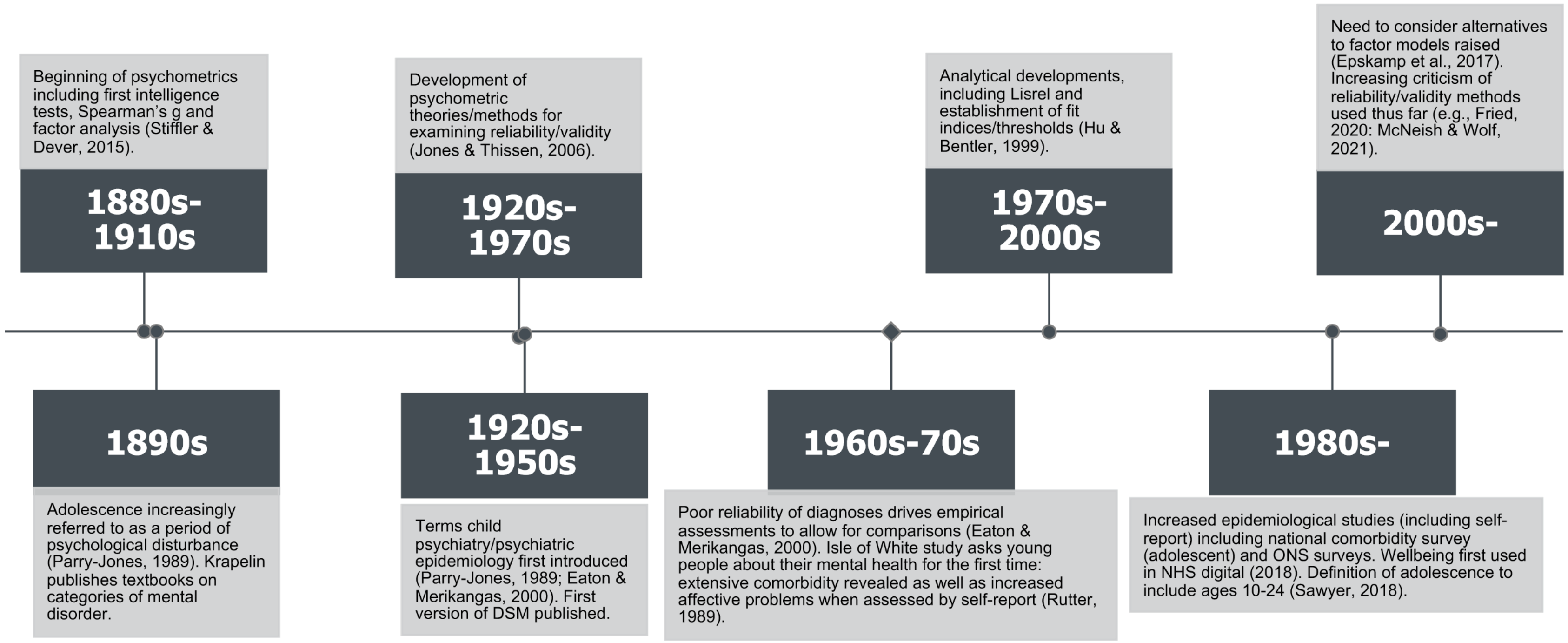
A Brief History of Adolescent Psychiatric Epidemiology and Implications for Current Practice

As noted above, adolescence is typically defined as a long transition between childhood and adulthood. Since puberty now occurs earlier, and adult roles are delayed compared to 150 years ago, it is now considered to span a longer period than previously, ages 10-24 (Sawyer et al., 2018). Nevertheless, adolescence as a discrete period of physical and social change has been consistently identified as important for mental health since the 1890s (Parry-Jones, 1989).

Though diagnostic categories and quantitative measurement of individual differences also date back as far as the late nineteenth century (Bentall, 2006; Jones & Thissen, 2006; Stiffler & Dever, 2015), large-scale estimation of population mental health did not begin until the second half of the twentieth century (Eaton & Merikangas, 2000). The intervening years saw key developments in psychometrics such as the development of classical test theory, and approaches to modeling validity and reliability (Jones & Thissen, 2006), though refinements in software and more robust standards (e.g., Hu & Bentler, 1999) were not achieved until much later in the twentieth century (Borsboom, 2006) and are ongoing (Epskamp, 2019; Epskamp et al., 2017; McNeish & Wolf, 2021). The limitations of theoretical inferences from psychometric models have also recently particularly gained attention (Fried, 2020a; van der Maas et al., 2006). Figure 1 provides a simplified timeline of psychometrics and adolescent psychiatric epidemiology since the late nineteenth century.

Figure 1.1

Timeline of Developments in Psychometrics and Adolescent Psychiatric Epidemiology



1880s-1910s

Beginning of psychometrics including first intelligence tests, Spearman's g and factor analysis (Stiffler & Dever, 2015).

1890s

Adolescence increasingly referred to as a period of psychological disturbance (Parry-Jones, 1989). Krapelin publishes textbooks on categories of mental disorder.

1920s-1970s

Development of psychometric theories/methods for examining reliability/validity (Jones & Thissen, 2006).

1920s-1950s

Terms child psychiatry/psychiatric epidemiology first introduced (Parry-Jones, 1989; Eaton & Merikangas, 2000). First version of DSM published.

1960s-70s

Poor reliability of diagnoses drives empirical assessments to allow for comparisons (Eaton & Merikangas, 2000). Isle of White study asks young people about their mental health for the first time: extensive comorbidity revealed as well as increased affective problems when assessed by self-report (Rutter, 1989).

1970s-2000s

Analytical developments, including Lisrel and establishment of fit indices/thresholds (Hu & Bentler, 1999).

1980s-

Increased epidemiological studies (including self-report) including national comorbidity survey (adolescent) and ONS surveys. Wellbeing first used in NHS digital (2018). Definition of adolescence to include ages 10-24 (Sawyer, 2018).

2000s-

Need to consider alternatives to factor models raised (Epskamp et al., 2017). Increasing criticism of reliability/validity methods used thus far (e.g., Fried, 2020: McNeish & Wolf, 2021).

As psychiatric epidemiology developed, it became clear that studying only those using services is problematic since not all those with relevant problems will access support (Switzer et al., 2013). This gap also clearly persists for adolescents today (Collishaw & Sellers, 2020). Large-scale quantitative evaluation of mental health in general population samples therefore became necessary to accommodate epidemiology's goals to estimate prevalence and understand etiology.

A landmark development in the understanding of young people's mental health came in the 1960s with the Isle of Wight studies (Rutter et al., 1976). These used a two-stage design, screening the entire population via parent and teacher questionnaires, before intensively interviewing those children who screened positive for disorders as well as their parents and teachers. This afforded a massive increase in sample size, and was the first time young people were directly interviewed as part of a systematic study (Rutter, 1989). The scope of the study and inclusion of young people's views revealed particular increases in internalizing problems, those associated with emotion, including depressive and anxiety symptoms and disorders, between early and mid-adolescence, compared to previous proxy-reported studies (Rutter, 1989).

The Isle of Wight studies' use of standardized measures, multiple informants and two-stage design went on to provide a blueprint for many later epidemiological studies with similar methods still often used (Polanczyk et al., 2015). While issues such as informant agreement have received increased attention since (for meta-analyses see Achenbach et al., 1987; De Los Reyes et al., 2015), it is worth considering the historical roots of informant choice: Until the 1960s (at the earliest), young people were not considered suitable informants of their own mental health, and even when these were used, elevated rates of internalizing problems compared to proxy-reports were questioned in terms of clinical significance (Rutter, 1989). Research was also hampered by issues such as the fact depressive problems were not considered to be an issue in young people by the many following psychodynamic theory (Rutter & Sroufe, 2000). Despite the known discrepancies between different psychiatrists' observations and emerging evidence that questionnaires could provide reliable information, clinical observation was seen as irreplaceable by some (Beck et al., 1961). Together these issues suggest a historical bias against hearing directly from young people which may have only gradually improved in the second half of the twentieth century.

It is possible this issue still persists to some extent. While some research points to severity being associated with self/proxy agreement, evidence is somewhat unclear about the relative validity of different informants, particularly when considering internalizing symptoms (De Los Reyes et al., 2015). In addition, though informant discrepancies are similar for adult mental health, this is not typically taken as evidence that adults cannot self-report (Achenbach et al., 2005). Given the value now placed on young peoples' perspectives (Deighton et al., 2014), and the recognition of internal experiences as important to development and mental health in adolescence (Rapee et al., 2019), it is likely that if methods were developed now from scratch, they might look quite different. Just as current views of mental health symptoms are inexorably tied to original frameworks (Bentall, 2006; Kendler, 2016), there may be a specific historical legacy for the question of whose view to trust in adolescent mental health assessment. Despite progress made through studies like the Isle of Wight, by 1980 few large-scale studies of psychiatric morbidity in young people shared standardized procedures, though by 2002 several standardized interviews were in use (Buka et al., 2002). Throughout this time period the possibility for comparison and wider inference therefore grew. These interviews consisted broadly of categorical approaches in which raters judged disorders to be present or absent (e.g., Shaffer et al., 2000), and scales based on factor analyses which typically represented broader syndromes such as internalizing/externalizing problems (e.g., Achenbach & Ruffle, 2000).

There are therefore measures that have been used for a relatively long time and to draw longitudinal and cross-cultural comparisons (e.g., NHS Digital, 2018; Polanczyk et al., 2015). However, given the typical lag between psychometric developments and their application to wider research (Borsboom, 2006), these have likely not been developed to today's standards. In this sense, their long standing is both a benefit and a risk. In addition, it has been argued that different measures will be more or less applicable in different contexts (Patalay & Fried, 2020), and common measurement is not sufficient to assume cohorts can be compared (Wicherts et al., 2004). Therefore, while consistently measured nationally representative data exists across decades (e.g., Pitchforth et al., 2019), this is not enough to assume meaningful trends can be robustly inferred (Collishaw, 2015).

For comparisons, stringent psychometric standards must be met, while available evidence suggests this is not the case for adolescent measures (Angold et al., 2012; Bentley et al., 2019; Reeves et al., 2016). While validation procedures may have been satisfactory 20 years ago (e.g., Goodman et al., 2000; Goodman et al., 1998), they are likely questionable by today's standards (Flake et al., 2017). The fact measures dating back this far are still commonly used (Deighton et al., 2019; NHS Digital, 2018; Patalay & Fitzsimons, 2018), is also a cause for concern since a seminal adolescent mental health paper from this time highlighted measurement as a problematic area that needed particular attention (Rutter & Sroufe, 2000). Robust validation is also vital when sum-score approaches are adopted, as is typically the case, since this imposes strict assumptions which should be checked (Borsboom, 2017; Fried et al., 2014; McNeish & Wolf, 2020). Furthermore, the use of measures to rate and categorise individuals, where underpinning scores are not reliable or valid, is considered unethical (Adams, 2000). The quality and fitness for purpose of available data must therefore be held to account.

Given these issues seem to be deeply embedded in existing adolescent mental health data, robust analytical strategies to accommodate, identify and improve problems are clearly needed. It has also been pointed out that large resource-intensive datasets are typically subject to pressures from different stake-holders and are therefore additionally vulnerable to compromises in what is included and analysed (Orben & Przybylski, 2019). There is clearly therefore a need for transparent, robust modeling of adolescent mental health when considering existing datasets and measures.

Ontological Problems with Mental Disorders and Their Measurement

Psychiatric models implicitly assume that distinct disorders exist and can therefore be measured. However, many have argued against this and these assumptions face considerable problems. For instance, objective (i.e. biological) tests for mental health problems have not emerged, nor do they currently seem appropriate, despite hopes from some for biomarkers or evidence from brain imaging (Timimi, 2014). In fact, in an attempt to overcome the disconnect between clinical classification and empirical studies, the Research Domain Criteria (RDoC) framework, which defined mental health problems as brain disorders, was proposed (Insel et al., 2010). While RDoC was hailed as an opportunity to improve the empirical evidence base of psychiatry and loosen the dominance of diagnostic manuals (Lilienfeld & Treadway, 2016), meta-analytic evidence suggests brain dysfunctions specific to syndromes or symptoms have not been found (Fried, 2020b; Li et al., 2020; Sprooten et al., 2017). The conceptualization and measurement of mental disorder has therefore been largely built on subjective expert opinion, with standardized measures judged against this.

However, the substantial disagreement between practitioners and diagnostic systems make such classification ill-suited to large-scale research (Williams et al., 1980). For instance, the prevalence of adolescent mental health problems has been found to be significantly moderated by diagnostic approach in meta-analyses with differences in criteria considered a probable reason (Bronsard et al., 2016; Polanczyk et al., 2015; Xu et al., 2018). This inconsistency has likely in turn contributed to the substantial heterogeneity in what is assessed by different questionnaires (Newson et al., 2020). In addition, it is important that assessment systems cover the entire population under study (Collishaw, 2015), which may not be the case when these are developed considering, and validated against the judgement of experts in impairment (Williams et al., 1980). Finally, while diagnostic systems have changed over the 100 or so years they have been in existence, they remain heavily influenced by original nosologies and are considered to be “historically contingent” (Kendler, 2016, p. 8); that is, they are dependent on the thinking of a few individuals and likely not objectively repeatable. There is some consensus that these dominant diagnostic frameworks have nevertheless been useful, at least to some extent (Borsboom, 2008), for instance to group those who share symptoms or provide diagnoses which in turn may trigger access to services. Nevertheless, together the issues described above suggest categorical diagnostic approaches are not well suited to epidemiological work.

To understand diagnostic issues further with specific relevance to adolescence and measurement, I provide examples of measure validation for the purpose of psychiatric epidemiology with young people. Specifically, I highlight issues in approaches used to estimate national prevalence in England (NHS Digital, 2018). The Development and Wellbeing Assessment (DAWBA) uses lay interviewers to ask parents and young people structured and semi-structured questions about criteria for certain DSM-IV/V and ICD-10 diagnoses, with varying coverage (Goodman et al., 2000). In addition to parent and young person interviews, teachers complete questionnaires. Interview and questionnaire data are then integrated to form a diagnosis, first by a computer and then by a clinical expert who considers the primary data as well as the computer decision. The final DAWBA outcome is therefore reduced to presence or absence of either a DSM or ICD-10 diagnosis or not otherwise specified (i.e. the clinician deemed clinically significant symptoms to be present without evidence that any specific diagnostic criteria were met). This format of the DAWBA is considered a ‘gold standard’ and is therefore used in high-stakes research such as estimating national prevalence (NHS Digital, 2018), though multi-categorical computer scoring is also available (Goodman et al., 2011).

The DAWBA is a standardized diagnostic interview, an approach which is designed to provide an efficient standardized approach to diagnosis for research (Reeves et al., 2016). As diagnostic tools, standardized through their relationship to classification systems, such interviews do not typically undergo extensive psychometric validation, with structural validity typically not considered (e.g., Piacentini et al., 1993). That is, validity is only considered in relation to the criterion, the diagnostic system itself, such that coverage of disorder criteria satisfies validity standards. This means underpinning disorders are assumed to be infallible. The process for validating diagnoses that is often cited proposes the following stages: clinical description, biological testing, specification of exclusion criteria and discrimination from other disorders, longitudinal and genetic studies (Robins & Guze, 1970). However, biological approaches never materialized (Andreasen, 1995), and comorbidity has become accepted as a rule (Borsboom et al., 2011; Lilienfeld, 2003). These stages have therefore proved somewhat limited. The validation of the interview *only* against diagnostic criteria therefore seems concerning. In contrast, psychometric development would typically entail qualitative and quantitative work to establish validity, such as coverage and accuracy (Slaney, 2017). Therefore, while validity in a psychometric sense consists of broader issues, such as how symptoms relate to one another and to other constructs (see Chapter 3), validity has only been considered in a very narrow sense based on mapping questions to criteria for standardized interviews. This is demonstrated below specifically for the DAWBA.

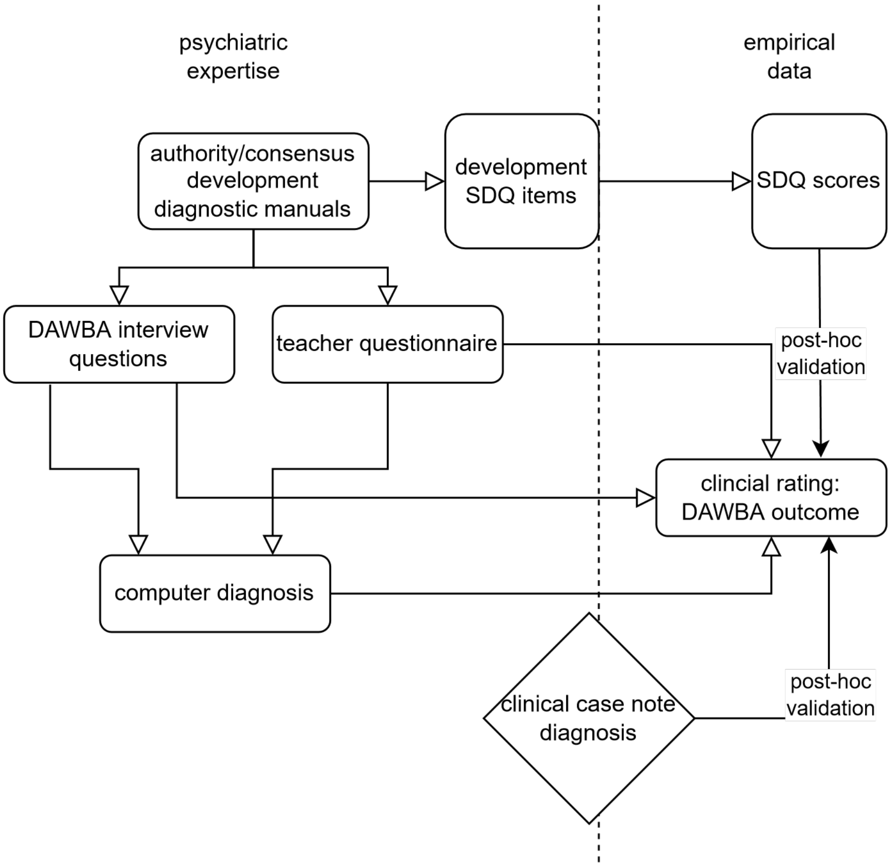
The development, validation and administration of the DAWBA is shown in Figure 2, alongside the development of the brief survey used to validate it (the strengths and difficulties questionnaire, SDQ; Goodman, 1997)[1]. Figure 2 demonstrates that the final DAWBA system is entirely developed in and validated against expert-derived psychiatric classification systems. While correspondence between raters (Ford et al., 2003), and against case-note diagnosis (Goodman et al., 2000) have been evaluated

[1] While the DAWBA has international versions, since psychometric properties are likely version dependent (Flake et al., 2017), I focus here on evidence for the English version.

posthoc, these checks all relate to diagnostic systems and expertise. Furthermore, rather than an iterative process of construct definition and measure refinement as would normally be the ideal in psychological measurement (Flake et al., 2017; Hughes, 2018; Slaney, 2017), this example includes no empirical data (e.g. cognitive interviews or pilot response patterns) to *inform* measure development. Omission of any of the work considering theory, empirical data and correspondence between the two, is considered a major threat to reliability and validity of findings (Rigdon et al., 2011). This is arguably particularly significant in this case because the nosology on which the DAWBA is based is itself controversial.

Figure 1.2

The Development of the DAWBA



Note. DAWBA = development and wellbeing assessment; SDQ = strengths and difficulties questionnaire.

In addition, two further problems are clear. First, the validity of the SDQ was subsequently judged on the basis of its relationship to the DAWBA. While this is not the only psychometric study underpinning the SDQ (Kersten et al., 2016), using each measure to validate the other could be problematic. Moreover, the SDQ has questionable psychometric properties itself (see Paper 3). Second, the type of data produced by systems such as the 'gold standard' version of the DAWBA (integrative best estimate systems which are commonplace in adolescent psychiatric epidemiology; Polanczyk et al., 2015), is arguably not consistent with current standards. From one perspective, the subjectivity of the rater allows consideration of issues such as different raters' relative comprehension to resolve discrepancies (Youth in Mind, 2017). However, on the other hand, to my knowledge, the computer algorithm is not reported, nor is it clear the extent to which raters should be or are influenced by this. A lack of transparency of this kind is considered a major threat to validity (Flake & Fried, 2020). The resulting binary data, diagnosis or no diagnosis, also represents a severe loss of information. Indeed, categorizing continuous data is known to cause a host of statistical problems in further analysis (Altman & Royston, 2006). Taken together, it seems clear such an approach, though considered a gold standard (NHS Digital, 2018; Reeves et al., 2016), faces serious limitations.

The above example demonstrates the ontological problem in mental health research described above. The only methods available to judge the presence or absence of disorders are clinical ratings, including systems such as the DAWBA. However, no evidence external to the nosological system itself is provided for the validity of such an approach. For instance, despite being a crucial requirement for the validation of a diagnostic measure, the authors acknowledge in the development of the DAWBA that they cannot estimate specificity and sensitivity:

> At least one ICD-10 or DSM-IV disorder was diagnosed in 11% of the community sample as compared with 92% of the clinic sample. This corresponds to a minimum estimate of 89% specificity in the community sample and 92% sensitivity in the clinic sample (based on the extreme and implausible assumption that all of the community sample with DAWBA diagnoses were false positives and all of the clinic sample without psychiatric diagnoses were false negatives) (Goodman et al., 2000, p. 649).

Similarly, Angold et al. (2012) in a comparison of three diagnostic interviews found significant discrepancies between assessments but concluded:

> Beginning with the DAWBA, when would one choose an interview that generated fewer, more severe cases? Two applications immediately spring to mind: services research and clinical trials. What use to tell policy makers that a third of all pediatric patients need psychiatric services? One in five is probably a more useful message (Angold et al., 2012, p. 515).

While pragmatic approaches will be needed, it seems problematic to base the validity and utility of a measure for estimating and understanding population prevalence on the palatability of rates it provides. This is particularly problematic in light of evidence that rates are changing over time (Collishaw, 2015).

Nevertheless, this is the state of the field in adolescent psychiatric epidemiology.

Kendler (2016) argues that a correspondence theory of truth, in which verification occurs through direct observation of the phenomenon, is unrealistic for psychiatry. Instead he suggests the more modest coherence theory of truth should be adopted in which verification can occur when models fit well with other things we clearly know. The validity of the DAWBA seems tied to a correspondence account, since it is developed against what are suggested to be true diagnostic systems. Since this is not verifiable, the system falls down. To move to coherence accounts of truth, it seems clear that more evidence should be gathered. For instance, what do key stakeholders (not just clinicians) perceive to be meaningful problems and how do identified indicators relate to one another empirically?

Such fixed approaches to diagnostic systems create additional problems. For instance, many diagnostic interviews, including DAWBA, employ a skip structure. These hold that if certain key symptoms are not experienced, it is not worth asking further questions since the criteria for diagnosis will not be met. However, it is not clear that the absence of criterion symptoms leads to the absence of others, or that this absence is meaningful for impairment, and so further information is lost (Fried et al., 2017). Similarly, working with data at the level of diagnosis loses information about subthreshold levels of the same experiences which could inform understanding (Ringwald et al., 2021). This issue is likely of particular interest in general population samples and prevention. In addition, a lack of validity beyond correspondence to a manual, seriously limits the applicability of research findings to practice (Jensen & Weisz, 2002). Given the lack of agreement between diagnostic interviews (Angold et al., 2012), and their incongruence with usual clinical practice (Reeves et al., 2016), serious questions arise about the value of using these to conduct epidemiological research (e.g., NHS Digital, 2018).

Alternatives to Clinical Rating and Mental Disorder

It is clear that a diagnostic approach to considering adolescent mental health in research faces limited validity. To improve understanding, three key approaches are adopted in the current thesis: First, positive and negative aspects of mental health are included to move away from diagnostic models and possibly better suit general population samples. Second, these aspects of mental health, symptoms, behaviours and thoughts, are considered at an item level, rather than assuming that they belong to a certain disorder or construct. Third, self-report data are used since this is considered to provide the best insight into internal experiences thought to be key to mental health in adolescence (Deighton et al., 2014; Rapee et al., 2019). Issues relating to these approaches are introduced in the following sections.

Existing Approaches Beyond Diagnoses

While diagnostic systems assume that mental health problems are categorical, many have argued that dimensional approaches should be adopted instead. Specifically, empirical investigation, using factor analysis for quantitative classification, has been argued to cut across expert consensus as a basis for organizing symptoms (Krueger et al., 2018). It is suggested that this addresses comorbidity since the associations between symptoms and syndromes are modelled. Indeed, there is a long history of such empirically-driven approaches in child and adolescent research (Achenbach, 1966; Achenbach & Edelbrock, 1978). However, a limitation of these approaches is that data-driven common factors can be difficult to interpret and there has been a tendency for researchers to overstate the substantive implications of empirically derived groupings (Littlefield et al., 2021; see also Chapter 3 and Paper 1).

Furthermore, while such models aim to move away from disorders, the factor models employed still make causal assumptions that latent dimensions cause symptoms (Fried, 2020a). Counter to this, many consider it likely that symptoms influence one another, which is not allowed for in restrictive factor models (Borsboom, 2017; Cramer et al., 2010). For instance, sleep deprivation may lead to poor concentration, rather than both being caused by an underlying internalizing factor. This problem, local independence, has been addressed by network models which explicitly model relationships between symptoms without latent variables. However, network models have tended to focus on the interaction of symptoms within a given disorder (Robinaugh, Hoekstra, et al., 2020). These approaches have therefore been limited to some extent through their link to problematic diagnostic nosology (discussed above).

In addition to considering how symptoms covary empirically, positive mental states might also be used to help improve understanding. While there has arguably been too strong a focus on disease models in adults (Kinderman, 2017), there has often been a more comprehensive approach with adolescents (e.g., Patalay & Fitzsimons, 2016; Suldo et al., 2016). This reflects theory that mental health should be considered a complete state, such that optimal functioning means being symptom free *and* feeling positive (Greenspoon & Saklofske, 2001; Keyes, 2005). This has been referred to as *complete mental health*, *the dual-factor approach*, or *the two continua model* (Greenspoon & Saklofske, 2001; Keyes, 2005; Westerhof & Keyes, 2010). Not only does this approach have intuitive appeal, it aligns with the World Health Organization's long-standing definition of mental health as comprising both absence of symptoms and positive functioning (WHO, 1946).

Despite this conceptual integrity, current empirical evidence is limited in several ways. First, the majority of research in this area has relied on norm-referenced or sample-based cut-offs, which by design create four groups for combinations of high/low symptoms/wellbeing: low symptoms and high wellbeing, symptomatic but content, high symptoms and low wellbeing, and low symptoms but low wellbeing (Moore et al., 2019). This is problematic as researchers have consistently inferred that simply the presence of these groups provides evidence of the dual-factor theory (e.g., Antaramian et al., 2010; Lyons et al., 2012). Furthermore, similar to diagnoses, as described above, reducing continuous data into categories can create statistical problems, including increased false positives and loss of information about variability within groups (Altman & Royston, 2006). Unlikely categories such as symptomatic but content, could also arise through invalid response patterns (Furlong et al., 2017).

Second, dual-factor studies assume wellbeing and mental health difficulties can be considered independent. Building a classification system based on multiple outcomes suggests each provides unique information which can be additively combined. For instance, the group symptomatic but content implies these constructs are conceptually and statistically unrelated. Statistically, this is unlikely since wellbeing and difficulties are typically assessed via self-report measures, meaning associations would be likely due to common method bias (Podsakoff et al., 2003). Also, conceptually, both constructs consider similar states, for instance both often include affect (Alexandrova & Haybron, 2016). Indeed, particularly strong relationships have been found between internalizing symptoms and wellbeing (up to r = -.68, Antaramian et al., 2010; Patalay & Fitzsimons, 2018; Suldo et al., 2011; The Children's Society, 2019). This problem that constructs of different names are assumed to measure different constructs is a recognized measurement issue, known as the jangle fallacy (Marsh, 1994). While the phenomenon of common indicators across mental health domains is well known in adult psychopathology (Borsboom et al., 2011), it has received little attention in the dual-factor literature. It is also likely that some aspects of wellbeing are directly related to diagnostic criteria for disorders (e.g., relaxedness; American Psychiatric Association, 2013).

A further problem is that much dual-factor research has tended to consider composites, with wellbeing often represented as a combination of positive affect, absence of negative affect, and life satisfaction scores, using separate scales for each, and different measures again to capture symptoms (Antaramian et al., 2010; Kelly et al., 2012; Lyons et al., 2013; Lyons et al., 2012; Suldo & Shaffer, 2008; Suldo et al., 2011; Suldo et al., 2016). Similarly, composites of internalizing and externalizing symptoms have been aggregated (Patalay & Fitzsimons, 2016), though these domains are typically considered distinct and not scored together (Goodman et al., 2010). This distinction between symptoms and wellbeing is not well defined, understood or evidenced. Without consideration of statistical and conceptual overlap, scales should not be combined simply because they are both labelled as assessing similar constructs, or kept separate solely because they were developed in different disciplines (Marsh, 1994). In addition, validated scales used in these studies, hold constant properties such as response format within but not between measures, and might introduce systematic error via, for instance, similar wording designed to maximize reliability within measures (Clifton, 2020). Together these issues make clear that underpinning measurement issues are complex and should be carefully considered before drawing substantive conclusions.

While there have been limitations in much of the dual-factor literature to date, the combined measurement of symptoms and wellbeing actually represents a potential psychometric opportunity. For instance, using indicators with diverse wording from different scales could maximize validity (Clifton, 2020). This should be a particular concern for questionnaire data with adolescents since it is known to be error-prone (Cornell et al., 2012; Cornell et al., 2014).

It has been pointed out that factor and network models both suffer from an over-reliance on data models, with too little attention to clearly defining theories (Fried, 2020a). Conceptual issues also seem to have received too little attention for models incorporating positive mental states. The current thesis therefore aimed to draw on the potential benefits of dimensional, network and positive approaches, while also considering conceptual issues alongside psychometric models.

Psychometric Background

While the current thesis aimed to utilise robust psychometric techniques to address the problems outlined thus far, the psychometric development and properties of measures used to collect data underpin any such work. Before considering constructs and modeling approaches (Chapters 2 and 3), a few contextual psychometric issues are therefore briefly introduced.

There has tended to be insufficient deployment of psychometric theory and methods in psychological science as a whole, hampering progress (Borsboom, 2006; Flake & Fried, 2020). Figure 1 demonstrates the lag between psychometric developments and their application in adolescent mental health specifically. For instance, while software and thresholds for factor structure quality were established between the 1970s-1990s, the field seems typified by measures that have undergone little structural analysis or rarely meet standards (Bentley et al., 2019). Needless to say, recent developments in psychometric theory such as network psychometrics (Epskamp et al., 2017) have not been accounted for in the development or deployment of measures.

This issue may have persisted for a number of reasons. First, the proper application of psychometric models to substantive areas is difficult both practically and technically (Borsboom, 2006). For instance, it can be difficult to publish as studies are seen as too technical for substantive journals and too applied for psychometric outlets (Borsboom, 2006). There are also policy pressures to provide rates of disorder (Costello, 2015), which as discussed is not readily conducive to robust methods. Psychologists typically also lack training in this area meaning fundamental misunderstandings often underpin measurement problems (Borsboom, 2006). On the one hand, these barriers mean there is often a lack of psychometric evidence beyond basics such as internal consistency (Bentley et al., 2019; Flake et al., 2017). On the other hand, the complexity of psychometrics is also often underestimated. For instance, a common misunderstanding is that good model fit can provide evidence for theoretical constructs (Fried, 2020a; Gignac & Kretzschmar, 2017).

The Need for Exploratory Psychometric Investigation

While psychometric methods are often referred to as confirmatory, e.g., confirmatory factor analysis, psychometric modeling cannot provide direct evidence for data generating models (van Bork et al., 2019). It has been argued that to address such problems, highly formalized mathematical models and simulations should be used to falsify precisely defined relationships (Fried, 2020a; Haslbeck et al., 2021; Robinaugh, Haslbeck, et al., 2020). While this may be intuitive for disorders like post-traumatic stress disorder which occurs in clearly defined populations and has relatively clear theoretical etiology, adolescent mental health problems are typified by disagreement in measurement (Angold et al., 2012; De Los Reyes et al., 2015), complex and heterogeneous antecedents (Cicchetti & Rogosch, 2002; Masten & Cicchetti, 2010; Rutter & Sroufe, 2000), and conflicting frameworks (see above, and Papers 1 and 4). I therefore argue that more exploratory work is needed since proceeding to confirmatory testing without sufficiently well-developed theory and measurement can be damaging to scientific progress (Scheel et al., 2020). Indeed, where phenomena and measurement are not well defined, as is the case for general mental health in adolescence, formal theory construction is not appropriate (Haslbeck et al., 2021).

The Need to Consider Self-Report Adolescent Mental Health Data

The sections above demonstrate a need to move away from diagnoses and employ more robust (exploratory) psychometric analysis. To address these needs, self-report data are needed for several reasons. First, though this is sometimes cited as more limited than clinician ratings (e.g., Benton et al., 2021), this cannot be considered to be the case when moving away from diagnoses: It is acknowledged that to understand symptoms and experiences, young people’s perspectives are needed, and that adolescents can validly and reliably report these (Riley, 2004). Whereas clinicians’ expertise can be required to determine whether symptoms constitute disorder, when this element is removed and analysis is conducted of individual symptoms and experiences, I argue there is no clear reason not to use self-report (further support for this is presented in the following paragraphs). While proxies are clearly necessary for younger children (Bell, 2007; de Leeuw, 2011), the importance of hearing from young people themselves weighs in favour of using self-report measures for adolescents, and they possess the relevant awareness and cognitive ability (Deighton et al., 2014; Riley, 2004).

Second, there are substantial discrepancies between different informants which are poorly understood. Meta-analyses have established low agreement between young people and adults (around r = .28 overall and .06 for diagnostic categorical approaches; Achenbach et al., 1987; De Los Reyes et al., 2015), suggesting they are not suitable for aggregation. These meta-analyses found much higher average agreement between proxy raters (e.g., parents and teachers), at around .60. This suggests that while each informant, for instance, self, parent and teacher, offer differing perspectives, self-report is particularly distinct. The agreement between self and proxy informants for mental health in adolescents is therefore sufficiently low to assume something different is measured for each.

Third, while there are methods that aim to find common variance, most likely to capture the construct, these are limited to a narrow range of conditions. These limited conditions include the fact informants should respond to the same measures, with properties such as response format held constant so that divergence is not introduced through measures (De Los Reyes et al., 2015). This is problematic since it is unusual for large studies, typically needed for psychometric modeling (Epskamp, 2020; Wang & Rhemtulla, 2021), to collect multiple informants for the same variable as limiting data burden is a major concern (Orben & Przybylski, 2019). In fact, self-report represents a reduced data burden compared to teachers and can be easier to obtain than parent reports (Humphrey & Wigelsworth, 2016). Also, while symptom approaches often have proxy informants (e.g., Goodman et al., 2010), this is rarer for positive approaches to mental health (Proctor et al., 2009; Tsang et al., 2012). In addition, only quite specific confirmatory factor analysis models can robustly capture interpretable method factors (Eid et al., 2016). Approaches estimating method effects are therefore limited to the assumptions of common causes and latent variables, which as discussed are not always appropriate.

De Los Reyes et al. (2013) argued that discrepancies are likely meaningful and should therefore be carefully considered and not merely partialled out as error. However, convincing approaches for doing so are not available. For instance, Makol et al. (2020) attempted one of the first empirical applications of modeling informant discrepancies. They used principal components analysis to distinguish theorized trait, context and perspective variance of different raters. While the authors argue this goes beyond classical test theory by considering some error to be measurable (i.e. systematic and not random), it nevertheless assumes that shared variance among raters gives access to the construct score. While a body of theory, the operations triad model (De Los Reyes et al., 2013), lies behind the modeling conducted by Makol and colleagues, just as an observed score cannot be equated with a construct, neither can a portion of variance reasonably be equated with a theorized context or perspective. This is a form of operationalism, equating a measurement with a theoretical construct (Borsboom, 2006). Furthermore, though Makol and colleagues expected the tri-partite model to offer different insight into social anxiety than would a composite score, they in fact found their trait score to be highly correlated with the composite score (r = .88).

In summary, issues with diagnoses and construct modeling for multiple informants mean self-report data are vital to further our understanding of adolescent mental health.

Considerations When Working with Self-Report Data

While the above section makes clear the need to draw on self-report data, this is by no means a perfect method. In fact, a recent systematic review suggested that the psychometric quality of self-report measures of adolescent general mental health is often unclear and low (Bentley et al., 2019). In addition, though the DAWBA example given above focuses on a diagnostic interview, diagnosis is also often used as the gold standard for item development and criterion validity testing for self-report measures, as was the case for symptom measures used in the empirical papers of the current thesis (Goodman, 2001; Patalay et al., 2014).

An additional consideration is the readability and age appropriateness of measures (de Leeuw, 2011; Patalay et al., 2018). Though even measures targeted at adults should be presented simply (Terwee et al., 2007), lack of consideration of this aspect is more likely to impact adolescents than their parents or teachers.

Another issue is the particular aspect of mental health reported on. Some evidence suggests internalizing symptoms should be reported on by adolescents themselves, while teacher and parent informants might be needed for externalizing behaviours (Humphrey & Wigelsworth, 2016). However, little incremental validity testing has been conducted when considering different informants for internalizing and externalizing symptoms (De Los Reyes et al., 2015). Furthermore, evidence of the relative predictive validity of different informants is analysed within a diagnostic framework. Given that external raters tend to agree with one another more than with the adolescent (De Los Reyes et al., 2015), it is perhaps unsurprising that proxy reports have been found to be more internally consistent (therefore according with theoretical nosology) and agree more with clinicians (Evans et al., 2020; Goodman et al., 2010). I argue self-report is still the best option despite these limitations given the above priorities. In addition, several other factors suggest self-report remains the best compromise. First, informant disagreement was consistently lower for internalizing than externalizing symptoms when meta-analysed (De Los Reyes et al., 2015). This suggests the problem of adults not having insight into adolescent's internalizing problems is a bigger issue than adolescents failing to report their externalizing behaviour. Second, externalizing symptoms have significant comorbidity with internalizing problems (Lilienfeld, 2003). This suggests that even when self-report externalizing symptom measures are not sensitive to problems, relevant information could still be picked up (at least for some young people) where general measures which also include internalizing symptoms are used. Third, given that item-level analysis has already been suggested to be necessary to overcome problems with diagnosis and categorization, theoretical or quantitative issues such as variability associated with individual items can be addressed on a case by case basis.

Summary of Current Barriers to Understanding Adolescent Mental Health

The above review of historical, psychometric and developmental issues highlights a number of problems that currently prevent high-quality research in adolescent mental health. First, categorical diagnoses are not falsifiable or empirically informed but rather have their roots in archaic systems. Second, these diagnoses in turn result in information loss by reducing continuous or ordinal data into fewer categories which is known to result in statistical biases. Third, the gold standard methods for making such diagnoses in large-scale research can lack transparent and robust development, and are therefore a threat to replicability and validity (Flake & Fried, 2020). Fourth, little attention has been paid to the psychometric underpinnings of the relationships between positive or negative mental health. Finally, the treatment of different informants' reports in analysis remains challenging. Given that mental health appears to be getting worse in adolescence despite decades of costly research (Collishaw, 2015), and the mental health of young people is thought to be more vulnerable than ever in the coming years given the corona virus pandemic (Benton et al., 2021; Han et al., 2020), these measurement problems must be addressed, before meaningful progress can be made.

The following priorities for the current thesis were therefore identified: the inclusion of approaches beyond disorder and diagnosis, self-report data, item-level analysis, and the need to consider validity more broadly than reference to psychiatric nosology.

References

Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs: General and Applied*, *80*(7), 1-37. https://doi.org/10.1037/h0093906

Achenbach, T. M., & Edelbrock, C. S. (1978). The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin*, *85*(6), 1275-1301. https://doi.org/10.1037/0033-2909.85.6.1275

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of Adult Psychopathology: Meta-Analyses and Implications of Cross-Informant Correlations. *Psychological Bulletin*, *131*(3), 361-382. https://doi.org/10.1037/0033-2909.131.3.361

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychol Bull*, *101*(2), 213-232.

Achenbach, T. M., & Ruffle, T. M. (2000). The Child Behavior Checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatrics in review*, *21*(8), 265-271.

Adams, K. M. (2000). Practical and ethical issues pertaining to test revisions. *Psychological assessment*, *12*(3), 281-286. https://doi.org/10.1037/1040-3590.12.3.281

Alexandrova, A., & Haybron, D. M. (2016). Is Construct Validation Valid? *Philosophy of Science*, *83*(5), 1098-1109. https://doi.org/10.1086/687941

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ*, *332*(7549), 1080. https://doi.org/10.1136/bmj.332.7549.1080

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5* (Fifth edition. ed.). American Psychiatric Association.

Andreasen, N. C. (1995). The validation of psychiatric diagnosis: new models and approaches. *The American journal of psychiatry*, *152*(2), 161-162. https://doi.org/10.1176/ajp.152.2.161

Angold, A., Erkanli, A., Copeland, W., Goodman, R., Fisher, P. W., & Costello, E. J. (2012). Psychiatric Diagnostic Interviews for Children and Adolescents: A Comparative Study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(5), 506-517. https://doi.org/https://doi.org/10.1016/j.jaac.2012.02.020

Antaramian, S. P., Huebner, S. E., Hills, K. J., & Valois, R. F. (2010). A Dual-Factor Model of Mental Health: Toward a More Comprehensive Understanding of Youth Functioning. *American Journal of Orthopsychiatry*, *80*(4), 462-472. https://doi.org/10.1111/j.1939-0025.2010.01049.x

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An Inventory for Measuring Depression. *Archives of General Psychiatry*, *4*(6), 561-571. https://doi.org/10.1001/archpsyc.1961.01710120031004

Bell, A. (2007). Designing and testing questionnaires for children. *Journal of Research in Nursing*, *12*(5), 461-469. https://doi.org/10.1177/1744987107079616

Bentall, R. (2006). Madness explained: Why we must reject the Kraepelinian paradigm and replace it with a ‘complaint-orientated’ approach to understanding mental illness. *Medical Hypotheses*, *66*(2), 220-233. https://doi.org/https://doi.org/10.1016/j.mehy.2005.09.026

Bentley, N., Hartley, S., & Bucci, S. (2019). Systematic Review of Self-Report Measures of General Mental Health and Wellbeing in Adolescent Mental Health. *Clinical Child and Family Psychology Review*, *22*(2), 225-252. https://doi.org/10.1007/s10567-018-00273-x

Benton, T. D., Boyd, R. C., & Njoroge, W. F. M. (2021). Addressing the Global Crisis of Child and Adolescent Mental Health. *JAMA Pediatrics*, *175*(11), 1108-1110. https://doi.org/10.1001/jamapediatrics.2021.2479

Blakemore, S.-J. (2019). Adolescence and mental health. *The Lancet*, *393*(10185), 2030-2031. https://doi.org/https://doi.org/10.1016/S0140-6736(19)31013-X

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*,*64*(9), 1089-1108. https://doi.org/https://doi.org/10.1002/jclp.20503

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5-13. https://doi.org/doi:10.1002/wps.20375

Borsboom, D., Cramer, A. O. J., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The Small World of Psychopathology. *PLOS ONE*, *6*(11), e27407. https://doi.org/10.1371/journal.pone.0027407

Bronsard, G., Alessandrini, M., Fond, G., Loundou, A., Auquier, P., Tordjman, S., & Boyer, L. (2016). The Prevalence of Mental Disorders Among Children and Adolescents in the Child Welfare System: A Systematic Review and Meta-Analysis. *Medicine*, *95*(7), e2622-e2622. https://doi.org/10.1097/MD.0000000000002622

Buka, S. L., Monuteaux, M., & Earlsi, F. (2002). The Epidemiology of Child and Adolescent Mental Disorders. In *Textbook in Psychiatric Epidemiology* (pp. 629-655). https://doi.org/10.1002/0471234311.ch23

Cicchetti, D., & Rogosch, F. A. (2002). A developmental psychopathology perspective on adolescence. *Journal of Consulting and Clinical Psychology*, *70*(1), 6-20. https://doi.org/10.1037/0022-006X.70.1.6

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, *25*(3), 259-270. https://doi.org/10.1037/met0000236

Collishaw, S. (2015). Annual Research Review: Secular trends in child and adolescent mental health. *Journal of Child Psychology and Psychiatry*, *56*(3), 370-393. https://doi.org/10.1111/jcpp.12372

Collishaw, S., & Sellers, R. (2020). Trends in Child and Adolescent Mental Health Prevalence, Outcomes, and Inequalities. In E. Taylor, F. Verhulst, J. C. M. Wong, & K. Yoshida (Eds.), *Mental Health and Illness of Children and Adolescents* (pp. 63-73). Springer Singapore. https://doi.org/10.1007/978-981-10-2348-4_9

Cornell, D., Klein, J., Konold, T., & Huang, F. (2012). Effects of validity screening items on adolescent survey data. *Psychological assessment*, *24*(1), 21-35. https://doi.org/10.1037/a0024824

Cornell, D. G., Lovegrove, P. J., & Baly, M. W. (2014). Invalid survey response patterns among middle school students. *Psychological assessment*, *26*(1), 277-287. https://doi.org/10.1037/a0034808

Costello, J. (2015). Commentary: ‘Diseases of the world’: from epidemiology to etiology of child and adolescent psychopathology – a commentary on Polanczyk et al. (2015). *Journal of ChildPsychology and Psychiatry*, *56*(3), 366-369. https://doi.org/10.1111/jcpp.12402

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, *33*(2-3), 137-150. https://doi.org/10.1017/S0140525X09991567

Dahl, R. E., Allen, N. B., Wilbrecht, L., & Suleiman, A. B. (2018). Importance of investing in adolescence from a developmental science perspective. *Nature*, *554*, 441. https://doi.org/10.1038/nature25770

de Leeuw, E. D. (2011). *Improving data quality when surveying children and adolescents: Cognitive and social development and its role in questionnaire construction and pretesting.* http://www.aka.fi/globalassets/awanhat/documents/tiedostot/lapset/presentations-of-the-annualseminar-10-12-may-2011/surveying-children-and-adolescents_de-leeuw.pdf

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The Validity of the Multi-Informant Approach to Assessing Child and Adolescent Mental Health. *Psychological Bulletin*, *141*(4), 858-900. https://doi.org/10.1037/a0038498

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. A. (2013). Principles Underlying the Use of Multiple Informants' Reports. *Annual Review of Clinical Psychology*, *9*(1), 123-149. https://doi.org/10.1146/annurev-clinpsy-050212-185617

Deighton, J., Croudace, T., Fonagy, P., Brown, J., Patalay, P., & Wolpert, M. (2014). Measuring mental health and wellbeing outcomes for children and adolescents to inform practice and policy: a review of child self-report measures. *Child and Adolescent Psychiatry and Mental Health*, *8*(1), 14. https://doi.org/10.1186/1753-2000-8-14

Deighton, J., Lereya, S. T., Casey, P., Patalay, P., Humphrey, N., & Wolpert, M. (2019). Prevalence of mental health problems in schools: poverty and other risk factors among 28 000 adolescents in England. *The British Journal of Psychiatry*, 1-3. https://doi.org/10.1192/bjp.2019.19

Eaton, W. W., & Merikangas, K. R. (2000). Psychiatric epidemiology: progress and prospects in the year 2000. *Epidemiologic reviews*, *22*(1), 29-34. https://doi.org/10.1093/oxfordjournals.epirev.a018022

Eid, M., Geiser, C., & Koch, T. (2016). Measuring Method Effects:From Traditional to Design-Oriented Approaches. *Current Directions in Psychological Science*, *25*(4), 275-280. https://doi.org/10.1177/0963721416649624

Epskamp, S. (2019). Reproducibility and Replicability in a Fast-Paced Methodological World. *Advances in Methods and Practices in Psychological Science*, *2*(2), 145-155. https://doi.org/10.1177/2515245919847421

Epskamp, S. (2020). Psychometric network models from time-series and panel data. *Psychometrika*. https://doi.org/10.1007/s11336-020-09697-3

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized Network Psychometrics: Combining Network and Latent Variable Models. *Psychometrika*, *82*(4), 904-927. https://doi.org/10.1007/s11336-017-9557-x

Evans, S. C., Bonadio, F. T., Bearman, S. K., Ugueto, A. M., Chorpita, B. F., & Weisz, J. R. (2020). Assessing the Irritable and Defiant Dimensions of Youth Oppositional Behavior Using CBCL and YSR Items. *Journal of Clinical Child & Adolescent Psychology*, *49*(6), 804-819. https://doi.org/10.1080/15374416.2019.1622119

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456-465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality*

Science, *8*(4), 370-378. https://doi.org/10.1177/1948550617693063

Ford, T., Goodman, R., & Meltzer, H. (2003). The British Child and Adolescent Mental Health Survey 1999: The Prevalence of DSM-IV Disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*(10), 1203-1211. https://doi.org/https://doi.org/10.1097/00004583-200310000-00011

Fried, E. I. (2020a). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, *31*(4), 271-288. https://doi.org/10.1080/1047840X.2020.1853461

Fried, E. I. (2020b). *Syllabus: On the Nature of Mental Illness.* Retrieved 23/02/2022 from https://osf.io/45unv/

Fried, E. I., Nesse, R. M., Zivin, K., Guille, C., & Sen, S. (2014). Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. *Psychological Medicine*, *44*(10), 2067-2076. https://doi.org/10.1017/S0033291713002900

Fried, E. I., van Borkulo, C. D., Cramer, A. O., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, *52*(1), 1-10. https://doi.org/10.1007/s00127-016-1319-z

Furlong, M. J., Fullchange, A., & Dowdy, E. (2017). Effects of mischievous responding on universal mental health screening: I love rum raisin ice cream, really I do! *School psychology quarterly : the official journal of the Division of School Psychology, American Psychological Association*, *32*(3), 320-335. https://doi.org/10.1037/spq0000168

Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence*, *62*, 138-147. https://doi.org/https://doi.org/10.1016/j.intell.2017.04.001

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to Use Broader Internalising and Externalising Subscales Instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British Parents, Teachers and Children. *Journal of*

Abnormal Child Psychology, *38*(8), 1179-1191. https://doi.org/10.1007/s10802-010-9434-x

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. *Journal of ChildPsychology and Psychiatry*, *38*(5), 581-586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Goodman, R. (2001). Psychometric Properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337-1345. https://doi.org/https://doi.org/10.1097/00004583-200111000-00015

Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The Development and Wellbeing Assessment: Description and Initial Validation of an Integrated Assessment of Child and Adolescent Psychopathology. *Journal of Child Psychology and Psychiatry*, *41*(5), 645-655. https://doi.org/10.1111/j.1469-7610.2000.tb02345.x

Goodman, A., Heiervang, E., Collishaw, S., & Goodman, R. (2011). The 'DAWBA bands' as an ordered-categorical measure of child mental health: description and validation in British and Norwegian samples. *Social Psychiatry and Psychiatric Epidemiology*, *46*(6), 521-532. https://doi.org/10.1007/s00127-010-0219-x

Goodman, R., Meltzer, H., & Bailey, V. (1998). The strengths and difficulties questionnaire: A pilot study on the validity of the self-report version. *European Child & Adolescent Psychiatry*, *7*(3), 125-130. https://doi.org/10.1007/s007870050057

Greenspoon, P. J., & Saklofske, D. H. (2001). Toward an Integration of Subjective Well-Being and Psychopathology. *Social Indicators Research*, *54*(1), 81-108. https://doi.org/10.1023/a:1007219227883

Han, R. H., Schmidt, M. N., Waits, W. M., Bell, A. K. C., & Miller, T. L. (2020). Planning for Mental Health Needs During COVID-19. *Current Psychiatry Reports*, *22*(12), 66. https://doi.org/10.1007/s11920-020-01189-6

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2021). Modeling psychopathology: From data models to formal theories. *Psychological Methods*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000303

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hughes, D. J. (2018). Psychometric Validity. In *The Wiley Handbook of Psychometric Testing* (pp. 751-779). https://doi.org/https://doi.org/10.1002/9781118489772.ch24

Humphrey, N., & Wigelsworth, M. (2016). Making the case for universal school-based mental health screening. *Emotional and Behavioural Difficulties*, *21*(1), 22-42. https://doi.org/10.1080/13632752.2015.1120051

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., . . . Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, *167*(7), 748-751. https://doi.org/10.1176/appi.ajp.2010.09091379

Jensen, A. L., & Weisz, J. R. (2002). Assessing match and mismatch between practitioner-generated and standardized interview-generated diagnoses for clinic-referred children and adolescents. *Journal of Consulting and Clinical Psychology*, *70*(1), 158-168. https://doi.org/10.1037/0022-006X.70.1.158

Jones, L. V., & Thissen, D. (2006). 1 A History and Overview of Psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 1-27). Elsevier. https://doi.org/https://doi.org/10.1016/S0169-7161(06)26001-2

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

Kelly, R. M., Hills, K. J., Huebner, E. S., & McQuillin, S. D. (2012). The Longitudinal Stability and Dynamics of Group Membership in the Dual-Factor Model of Mental Health: Psychosocial

Predictors of Mental Health. *Canadian Journal of School Psychology*, *27*(4), 337-355. https://doi.org/10.1177/0829573512458505

Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry*, *15*(1), 5-12. https://doi.org/10.1002/wps.20292

Kersten, P., Czuba, K., McPherson, K., Dudley, M., Elder, H., Tauroa, R., & Vandal, A. (2016). A systematic review of evidence for the psychometric properties of the Strengths and Difficulties Questionnaire. *International Journal of Behavioral Development*, *40*(1), 64-75. https://doi.org/10.1177/0165025415570647

Keyes, C. L. M. (2005). Mental Illness and/or Mental Health? Investigating Axioms of the Complete State Model of Health. *Journal of Consulting and Clinical Psychology*, *73*(3), 539-548. https://doi.org/10.1037/0022-006X.73.3.539

Kinderman, P. (2017). A Manifesto for Psychological Health and Wellbeing. In J. Davies (Ed.), *The Sedated Society: The Causes and Harms of our Psychiatric Drug Epidemic* (pp. 271-301). Springer International Publishing. https://doi.org/10.1007/978-3-319-44911-1_11

Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., . . . Zimmermann, J. (2018). Progress in achieving quantitative classification of psychopathology. *World Psychiatry*, *17*(3), 282-293. https://doi.org/doi:10.1002/wps.20566

Li, T., Wang, L., Camilleri, J. A., Chen, X., Li, S., Stewart, J. L., . . . Feng, C. (2020). Mapping common grey matter volume deviation across child and adolescent psychiatric disorders. *Neuroscience & Biobehavioral Reviews*, *115*, 273-284. https://doi.org/https://doi.org/10.1016/j.neubiorev.2020.05.015

Lilienfeld, S. O. (2003). Comorbidity Between and Within Childhood Externalizing and Internalizing Disorders: Reflections and Directions. *Journal of Abnormal Child Psychology*, *31*(3), 285-291. https://doi.org/10.1023/a:1023229529866

Lilienfeld, S. O., & Treadway, M. T. (2016). Clashing Diagnostic Approaches: DSM-ICD Versus RDoC. *Annual Review of Clinical Psychology*, *12*(1), 435-463. https://doi.org/10.1146/annurev-clinpsy-

021815-093122

Littlefield, A. K., Lane, S. P., Gette, J. A., Watts, A. L., & Sher, K. J. (2021). The “Big Everything”: Integrating and investigating dimensional models of psychopathology, personality, personality pathology, and cognitive functioning. *Personality Disorders: Theory, Research, and Treatment*,*12*(2), 103-114. https://doi.org/10.1037/per0000457

Lyons, M. D., Huebner, E. S., & Hills, K. J. (2013). The Dual-Factor Model of Mental Health: A Short-Term Longitudinal Study of School-Related Outcomes. *Social Indicators Research*, *114*(2), 549-565. https://doi.org/10.1007/s11205-012-0161-2

Lyons, M. D., Huebner, E. S., Hills, K. J., & Shinkareva, S. V. (2012). The Dual-Factor Model of Mental Health: Further Study of the Determinants of Group Differences. *Canadian Journal of School Psychology*, *27*(2), 183-196. https://doi.org/10.1177/0829573512443669

Makol, B. A., Youngstrom, E. A., Racz, S. J., Qasmieh, N., Glenn, L. E., & De Los Reyes, A. (2020). Integrating Multiple Informants’ Reports: How Conceptual and Measurement Models May Address Long-Standing Problems in Clinical Decision-Making. *Clinical Psychological Science*, *0*(0), 2167702620924439. https://doi.org/10.1177/2167702620924439

Marsh, H. W. (1994). Sport Motivation Orientations: Beware of Jingle-Jangle Fallacies. *Journal of Sport and Exercise Psychology*, *16*(4), 365-380. https://doi.org/10.1123/jsep.16.4.365

Masten, A. S., & Cicchetti, D. (2010). Developmental cascades. *Development and Psychopathology*, *22*(3), 491-495. https://doi.org/10.1017/S0954579410000222

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01398-0

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000425

Merikangas, K. R., He, J.-p., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., . . . Swendsen, J. (2010). Lifetime Prevalence of Mental Disorders in U.S. Adolescents: Results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A). *Journal of the AmericanAcademy of Child & Adolescent Psychiatry*, *49*(10), 980-989. https://doi.org/https://doi.org/10.1016/j.jaac.2010.05.017

Moore, S. A., Dowdy, E., Nylund-Gibson, K., & Furlong, M. J. (2019). A latent transition analysis of the longitudinal stability of dual-factor mental health in adolescence. *Journal of School Psychology*, *73*, 56-73. https://doi.org/https://doi.org/10.1016/j.jsp.2019.03.003

Newson, J. J., Hunter, D., & Thiagarajan, T. C. (2020). The Heterogeneity of Mental Health Assessment. *Frontiers in Psychiatry*, *11*(76). https://doi.org/10.3389/fpsyt.2020.00076

NHS Digital. (2018). *Mental Health of Children and Young People in England, 2017 Summary of key findings*. https://files.digital.nhs.uk/F6/A5706C/MHCYP%202017%20Summary.pdf

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*(2), 173-182. https://doi.org/10.1038/s41562-018-0506-1

Parry-Jones, W. L. (1989). The History of Child and Adolescent Psychiatry: Its Present Day Relevance. *Journal of Child Psychology and Psychiatry*, *30*(1), 3-11. https://doi.org/10.1111/j.1469-7610.1989.tb00766.x

Patalay, P., Deighton, J., Fonagy, P., Vostanis, P., & Wolpert, M. (2014). Clinical validity of the Me and My School questionnaire: a self-report mental health measure for children and adolescents. *Child and Adolescent Psychiatry and Mental Health*, *8*(1), 17. https://doi.org/10.1186/1753-2000-8-17

Patalay, P., & Fitzsimons, E. (2016). Correlates of Mental Illness and Wellbeing in Children: Are They the Same? Results From the UK Millennium Cohort Study. *J Am Acad Child Adolesc Psychiatry*, *55*(9), 771-783. https://doi.org/10.1016/j.jaac.2016.05.019

Patalay, P., & Fitzsimons, E. (2018). Development and predictors of mental ill-health and wellbeing from childhood to adolescence. *Social Psychiatry and Psychiatric Epidemiology*, *53*, 1311–1323 https://doi.org/10.1007/s00127-018-1604-0

Patalay, P., & Fried, E. I. (2020). Editorial Perspective: Prescribing measures: unintended negative consequences of mandating standardized mental health measurement. *Journal of Child Psychology and Psychiatry*, *62*(8). https://doi.org/10.1111/jcpp.13333

Patalay, P., Hayes, D., & Wolpert, M. (2018). Assessing the readability of the self-reported Strengths and Difficulties Questionnaire. *BJPsych Open*, *4*(2), 55-57. https://doi.org/10.1192/bjo.2017.13

Patton, G. C., Sawyer, S. M., Santelli, J. S., Ross, D. A., Afifi, R., Allen, N. B., . . . Viner, R. M. (2016). Our future: a Lancet commission on adolescent health and wellbeing. *The Lancet*, *387*(10036), 2423-2478. https://doi.org/https://doi.org/10.1016/S0140-6736(16)00579-1

Piacentini, J., Shaffer, D., Fisher, P., Schwab-Stone, M., Davies, M., & Gioia, P. (1993). The Diagnostic Interview Schedule for Children-Revised Version (DISC-R): III. Concurrent Criterion Validity. *Journal of the American Academy of Child & Adolescent Psychiatry*, *32*(3), 658-665. https://doi.org/https://doi.org/10.1097/00004583-199305000-00025

Pitchforth, J., Fahy, K., Ford, T., Wolpert, M., Viner, R. M., & Hargreaves, D. S. (2019). Mental health and well-being trends among children and young people in the UK, 1995–2014: analysis of repeated cross-sectional national health surveys. *Psychological Medicine*, *49*(8), 1275-1285. https://doi.org/10.1017/S0033291718001757

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879-903. https://doi.org/10.1037/0021-9010.88.5.879

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

Proctor, C., Linley, P. A., & Maltby, J. (2009). Youth life satisfaction measures: A review. *The Journal of Positive Psychology*, *4*(2), 128-144. https://doi.org/http://dx.doi.org/10.1080/17439760802650816

Rapee, R. M., Oar, E. L., Johnco, C. J., Forbes, M. K., Fardouly, J., Magson, N. R., & Richardson, C. E. (2019). Adolescent development and risk for the onset of social-emotional disorders: A review and conceptual model. *Behaviour Research and Therapy*, *123*, 103501. https://doi.org/https://doi.org/10.1016/j.brat.2019.103501

Reeves, K., Charter, E., & Ford, T. (2016). Measurement Issues: Is standardised diagnostic assessment feasible as an adjunct to clinical practice? A systematic review. *Child and Adolescent Mental Health*, *21*(1), 51-63. https://doi.org/https://doi.org/10.1111/camh.12089

Rigdon, E. E., Preacher, K. J., Lee, N., Howell, R. D., Franke, G. R., & Borsboom, D. (2011). Avoiding measurement dogma: A response to Rossiter. *European Journal of Marketing*, *45*(11-12), 1589-1600. https://doi.org/10.1108/03090561111167306

Riley, A. W. (2004). Evidence That School-Age Children Can Self-Report on Their Health. *Ambulatory Pediatrics*, *4*(4), 371-376. https://doi.org/https://doi.org/10.1367/A03-178R.1

Ringwald, W. R., Forbes, M. K., & Wright, A. G. C. (2021). Meta-analysis of structural evidence for the Hierarchical Taxonomy of Psychopathology (HiTOP) model. *Psychological Medicine*, 1-14. https://doi.org/10.1017/S0033291721001902

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. (2020). Invisible Hands and Fine Calipers: A Call to Use Formal Theory as a Toolkit for Theory Construction. *PsyArXiv*. https://doi.org/10.31234/osf.io/ugz7y

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R., & Borsboom, D. (2020). The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*, *50*(3), 353-366. https://doi.org/10.1017/S0033291719003404

Robins, E., & Guze, S. B. (1970). Establishment of Diagnostic Validity in Psychiatric Illness: Its Application to Schizophrenia. *American Journal of Psychiatry*, *126*(7), 983-987. https://doi.org/10.1176/ajp.126.7.983

Rutter, M. (1989). Isle of Wight Revisited: Twenty-five Years of Child Psychiatric Epidemiology. *Journal of the American Academy of Child & Adolescent Psychiatry*, *28*(5), 633-653. https://doi.org/https://doi.org/10.1097/00004583-198909000-00001

Rutter, M., & Sroufe, L. A. (2000). Developmental psychopathology: Concepts and challenges.*Development and Psychopathology*, *12*(3), 265-296. https://doi.org/10.1017/S0954579400003023

Rutter, M., Tizard, J., Yule, W., Graham, P., & Whitmore, K. (1976). Isle of Wight Studies, 1964–1974. *Psychological Medicine*, *6*(2), 313-332. https://doi.org/10.1017/S003329170001388X

Sawyer, S. M., Azzopardi, P. S., Wickremarathne, D., & Patton, G. C. (2018). The age of adolescence. *The Lancet Child & Adolescent Health*, *2*(3), 223-228. https://doi.org/10.1016/S2352-4642(18)30022-1

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, *16*(4), 744-755. https://doi.org/10.1177/1745691620966795

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC-IV): Description, Differences From Previous Versions, and Reliability of Some Common Diagnoses. *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(1), 28-38. https://doi.org/https://doi.org/10.1097/00004583-200001000-00014

Slaney, K. (2017). Construct Validity: Developments and Debates. In *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions* (pp. 83-109). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-38523-9_4

Solmi, M., Radua, J., Olivola, M., Croce, E., Soardo, L., Salazar de Pablo, G., . . . Fusar-Poli, P. (2021). Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-021-01161-7

Sprooten, E., Rasgon, A., Goodman, M., Carlin, A., Leibu, E., Lee, W. H., & Frangou, S. (2017). Addressing reverse inference in psychiatric neuroimaging: Meta-analyses of task-related brain activation in common mental disorders. *Human Brain Mapping*, *38*(4), 1846-1864.https://doi.org/https://doi.org/10.1002/hbm.23486

Stiffler, M. C., & Dever, B. V. (2015). History of Screening Practices, Mental Health Assessment, and Classification in the USA. In *Mental Health Screening at School: Instrumentation, Implementation, and Critical Issues* (pp. 5-26). Springer International Publishing. https://doi.org/10.1007/978-3-319-19171-3_2

Suldo, S., & Shaffer, E. J. (2008). Looking Beyond Psychopathology: The Dual-Factor Model of Mental Health in Youth. *School Psychology Review*, *37*(1), 52-68.

Suldo, S., Thalji, A., & Ferron, J. (2011). Longitudinal academic outcomes predicted by early adolescents' subjective well-being, psychopathology, and mental health status yielded from a dual factor model. *The Journal of Positive Psychology*, *6*(1), 17-30. https://doi.org/10.1080/17439760.2010.536774

Suldo, S., Thalji-Raitano, A., Kiefer, S. M., & Ferron, J. M. (2016). Conceptualizing High School Students' Mental Health Through a Dual-Factor Model. *School Psychology Review*, *45*(4), 434-457. https://doi.org/10.17105/spr45-4.434-457

Switzer, G. E., Dew, M. A., & Bromet, E. J. (2013). Issues in Mental Health Assessment. In C. S. Aneshensel, J. C. Phelan, & A. Bierman (Eds.), *Handbook of the Sociology of Mental Health* (pp. 115-141). Springer Netherlands. https://doi.org/10.1007/978-94-007-4276-5_7

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., . . . de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34-42. https://doi.org/https://doi.org/10.1016/j.jclinepi.2006.03.012

The Children's Society. (2019). *The Good Childhood Report 2019*. https://www.childrenssociety.org.uk/sites/default/files/the_good_childhood_report_2019.pdf

Timimi, S. (2014). No more psychiatric labels: Why formal psychiatric diagnostic systems should be abolished. *International Journal of Clinical and Health Psychology*, *14*(3), 208-215. https://doi.org/https://doi.org/10.1016/j.ijchp.2014.03.004

Tsang, K. L. V., Wong, P. Y. H., & Lo, S. K. (2012). Assessing psychosocial well-being of adolescents: a systematic review of measuring instruments. *Child: Care, Health and Development*, *38*(5), 629-646. https://doi.org/https://doi.org/10.1111/j.1365-2214.2011.01355.x

van Bork, R., Rhemtulla, M., Waldorp, L. J., Kruis, J., Rezvanifar, S., & Borsboom, D. (2019). Latent Variable Models and Networks: Statistical Equivalence and Testability. *Multivariate Behavioral Research*, 1-24. https://doi.org/10.1080/00273171.2019.1672515 van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., &

Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842-861. https://doi.org/10.1037/0033-295X.113.4.842

Wang, Y. A., & Rhemtulla, M. (2021). Power Analysis for Parameter Estimation in Structural Equation Modeling: A Discussion and Tutorial. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920918253. https://doi.org/10.1177/2515245920918253

Westerhof, G. J., & Keyes, C. L. M. (2010). Mental Illness and Mental Health: The Two Continua Model Across the Lifespan. *Journal of Adult Development*, *17*(2), 110-119. https://doi.org/10.1007/s10804-009-9082-y

WHO. (1946). Constitution *of the World Health Organization*. http://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf?ua=1

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*(5), 509-537. https://doi.org/https://doi.org/10.1016/j.intell.2004.07.002

Williams, P., Tarnopolsky, A., & Hand, D. (1980). Case definition and case identification in psychiatric epidemiology: review and assessment. *Psychological Medicine*, *10*(1), 101-114. https://doi.org/10.1017/S0033291700039635

Xu, D.-D., Rao, W.-W., Cao, X.-L., Wen, S.-Y., Che, W.-I., Ng, C. H., . . . Xiang, Y.-T. (2018). Prevalence of major depressive disorder in children and adolescents in China: A systematic review and metaanalysis. *Journal of Affective Disorders*, *241*, 592-598. https://doi.org/https://doi.org/10.1016/j.jad.2018.07.083

Youth in Mind. (2017). *Clinical rating: the human expertise at the heart of the system*. Retrieved 05/02/2022 from https://dawba.info/d0.html

Chapter 2: Positive and Negative Mental Health: Constructs and Indicators

Chapter 1 made clear a need to move beyond diagnoses, drawing instead on developments from dimensional, network, and positive mental health approaches. The lack of conceptual clarity between positive and negative approaches was also introduced, and this is further explored in the papers of the thesis. The aim of this chapter is therefore not to provide a definitive account of constructs included in the papers of the thesis, since it is likely these are not cleanly demarcated. This issue has not been extensively considered but some literature has already highlighted that both symptoms and wellbeing can be measured for similar purposes and contexts (Alexandrova & Haybron, 2016; Bartels et al., 2013; Bentley et al., 2019; Deighton et al., 2014). Similarly, the empirical relationship between internalizing and externalizing problems has been highlighted (Achenbach et al., 2016; Lilienfeld, 2003). Instead, broad parameters that informed decisions in the papers of the current thesis are set out. The final paper of the thesis in fact systematically analyses the content of measures and constructs in adolescent general mental health since this was an identified gap. Paper 4 therefore builds on the basic rubric set out here (which informs particularly the first three papers).

Constructs, Domains, and Indicators

General mental health is a framework used in the current thesis to accommodate positive and negative states. This echoes two recent reviews of self-report measures for adolescent mental health (Bentley et al., 2019; Deighton et al., 2014). In the current thesis, these positive and negative states are not considered categorically via diagnoses or classes such as flourishing (Keyes, 2005), given the statistical problems of reducing information, and problems of conflicting frameworks that might overlap. General mental health is therefore considered a constellation of feelings, behaviors, thoughts, and experiences. In order to guide the reader and contextualize the thesis within the broader literature, a few key parameters are defined here.

First, the terms *construct* and *domain* are used interchangeably in the papers of the thesis to refer to a theoretical unobserved component within general mental health such as internalizing symptoms or wellbeing. These components are observed through items and represent groups of symptoms or experiences that are considered to cluster together (typically based on a combination of statistical and theoretical models). Each construct also typically should show discriminant validity from others, based on theory. However, the critical treatment of constructs is a fundamental strand of the current thesis, building on work that has pointed out theoretical and statistical problems with the treatment of constructs in similar literature. For instance, Fried (2017) demonstrated that the construct of depression is inconsistently defined based on content analysis of typically used measures. Similarly, the constructs affect, wellbeing and depression are likely all consistent with experiences such as happiness/sadness or relaxation (Alexandrova & Haybron, 2016).

This makes clear that units *within* constructs also need to be considered. The overlap of symptoms within disorders is well known (Borsboom et al., 2011) but has not been considered in detail in non-diagnostic approaches to mental health. To facilitate this, the term *indicator* is used to refer to a feeling, thought, experience, symptom, or behaviour which is included in a given construct. This is typically captured by a single item, though some items may capture more than one indicator and different scales can capture the same indicator via multiple items. These units are therefore indicators of constructs (as in latent variable models) or within a complex system (network models).

The following sections briefly introduce constructs and the indicators which are typically considered within them when adolescents are asked to self-report on general mental health.

Symptoms and Mental Ill Health

Self-report general symptom measures have tended to focus on the most common problems, i.e., internalizing and externalizing problems (Deighton et al., 2013). While internalizing difficulties typically include sadness, anxiety, and somatic indicator types, externalizing difficulties tend to include indicators of conduct and hyperactivity-inattention problems, such as concentration and rule-breaking (Achenbach et al., 2016). This means other symptom types including eating, thought, or personality problems are typically excluded. This means sensitivity for common problems may be improved, while rarer experiences are missed. Indicators of mental ill-health in general measures, as considered in the current thesis, therefore incorporate a range of symptoms. However, at the construct level, these are inconsistently defined and there is likely overlap which explains some comorbidity (Achenbach et al., 2016). For instance, concentration problems could indicate inattentive or depressive disorders (American Psychiatric Association, 2013). Careful consideration of the theoretical content of items is therefore needed before modeling indicators and constructs.

Wellbeing

Similar problems with demarcation are evident for wellbeing. Various domains of wellbeing have been defined as individual constructs, such as subjective wellbeing, life satisfaction, and eudaimonic wellbeing, though the term is also sometimes used to refer to mental ill health (e.g., Fuhrmann et al., 2021). The subjective wellbeing model focuses on hedonic wellbeing and consists of life satisfaction and affect (Ryan & Deci, 2001). Eudaimonic wellbeing seeks to consider the construct beyond straightforward happiness, for instance via components such as autonomy and social relationships (Ryff & Keyes, 1995). However, eudaimonia is poorly defined, consisting of diffuse theoretical models which lack a unified approach to measurement, and it overlaps conceptually and empirically with hedonic wellbeing (Disabato et al., 2016; Kashdan et al., 2008).

Summary: Inclusion Criteria to be Considered as General Mental Health in the Current Thesis

To be considered constructs or indicators of general mental ill health in the current thesis, items and subscales had to relate to distress or be considered indicative of psychopathology (i.e., be included in a diagnostic system). The consideration of symptoms is therefore somewhat defined by diagnostic systems, though diagnoses were not used categorically. Wellbeing was defined in the current thesis as positive mental states, which included happiness and eudaimonic wellbeing. However, for both symptoms and wellbeing, indicators and constructs were excluded if it was likely they were part of proximal domains, e.g., autonomy, which could be considered antecedents, outcomes, or resilience factors (Fritz et al., 2018; Kashdan et al., 2008). For instance, the prosocial and peer problems subscales of the SDQ were mostly not considered to be part of mental health (Papers 2 and 4), given their focus on social skills, and consistent with work by others (Patalay & Fitzsimons, 2016)[2]. A transparent approach to selecting and coding items was adopted across the thesis (for instance the extensive supplementary material provided for Papers 2 and 4), given that some subjectivity was involved in this. Nevertheless, given the lack of

[2] An exception is one of the peer problems items in Paper 2, “I am usually on my own. I generally play alone or keep to myself”, which was considered to be consistent with internalizing problems.

conceptual clarity within and between positive and negative approaches to mental health in adolescence (see Chapter 1 and Paper 4), an attempt was made to make decisions simple and transparent.

References

Achenbach, T. M., Ivanova, M. Y., Rescorla, L. A., Turner, L. V., & Althoff, R. R. (2016). Internalizing/Externalizing Problems: Review and Recommendations for Clinical and Research Applications. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(8), 647-656. https://doi.org/https://doi.org/10.1016/j.jaac.2016.05.012

Alexandrova, A., & Haybron, D. M. (2016). Is Construct Validation Valid? *Philosophy of Science*, *83*(5), 1098-1109. https://doi.org/10.1086/687941

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5* (Fifth edition. ed.). American Psychiatric Association.

Bartels, M., Cacioppo, J. T., van Beijsterveldt, T. C. E. M., & Boomsma, D. I. (2013). Exploring the Association Between Well-Being and Psychopathology in Adolescents. *Behavior Genetics*, *43*(3), 177-190. https://doi.org/10.1007/s10519-013-9589-7

Bentley, N., Hartley, S., & Bucci, S. (2019). Systematic Review of Self-Report Measures of General Mental Health and Wellbeing in Adolescent Mental Health. *Clinical Child and Family PsychologyReview*, *22*(2), 225-252. https://doi.org/10.1007/s10567-018-00273-x

Borsboom, D., Cramer, A. O. J., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The Small World of Psychopathology. *PLOS ONE*, *6*(11), e27407. https://doi.org/10.1371/journal.pone.0027407

Deighton, J., Croudace, T., Fonagy, P., Brown, J., Patalay, P., & Wolpert, M. (2014). Measuring mental health and wellbeing outcomes for children and adolescents to inform practice and policy: a review of child self-report measures. *Child and Adolescent Psychiatry and Mental Health*, *8*(1), 14. https://doi.org/10.1186/1753-2000-8-14

Deighton, J., Tymms, P., Vostanis, P., Belsky, J., Fonagy, P., Brown, A., . . . Wolpert, M. (2013). The Development of a School-Based Measure of Child Mental Health. *Journal of Psychoeducational Assessment*, *31*(3), 247-257. https://doi.org/10.1177/0734282912465570

Disabato, D. J., Goodman, F. R., Kashdan, T. B., Short, J. L., & Jarden, A. (2016). Different types of wellbeing? A cross-cultural examination of hedonic and eudaimonic well-being. *Psychological assessment*, *28*(5), 471-482. https://doi.org/10.1037/pas0000209

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191-197.https://doi.org/https://doi.org/10.1016/j.jad.2016.10.019

Fritz, J., de Graaff, A. M., Caisley, H., van Harmelen, A.-L., & Wilkinson, P. O. (2018). A Systematic Review of Amenable Resilience Factors That Moderate and/or Mediate the Relationship Between Childhood Adversity and Mental Health in Young People. *Frontiers in Psychiatry*, *9*(230). https://doi.org/10.3389/fpsyt.2018.00230

Fuhrmann, D., van Harmelen, A.-L., & Kievit, R. A. (2021). Well-Being and Cognition Are Coupled During Development: A Preregistered Longitudinal Study of 1,136 Children and Adolescents. *Clinical Psychological Science*, *0*(0), 21677026211030211. https://doi.org/10.1177/21677026211030211

Kashdan, T. B., Biswas-Diener, R., & King, L. A. (2008). Reconsidering happiness: the costs of distinguishing between hedonics and eudaimonia. *The Journal of Positive Psychology*, *3*(4), 219-233. https://doi.org/10.1080/17439760802303044

Keyes, C. L. M. (2005). Mental Illness and/or Mental Health? Investigating Axioms of the Complete State Model of Health. *Journal of Consulting and Clinical Psychology*, *73*(3), 539-548. https://doi.org/10.1037/0022-006X.73.3.539

Lilienfeld, S. O. (2003). Comorbidity Between and Within Childhood Externalizing and Internalizing Disorders: Reflections and Directions. *Journal of Abnormal Child Psychology*, *31*(3), 285-291. https://doi.org/10.1023/a:1023229529866

Patalay, P., & Fitzsimons, E. (2016). Correlates of Mental Illness and Wellbeing in Children: Are They the Same? Results From the UK Millennium Cohort Study. *J Am Acad Child Adolesc Psychiatry*, *55*(9), 771-783. https://doi.org/10.1016/j.jaac.2016.05.019

Ryan, R. M., & Deci, E. L. (2001). On Happiness and Human Potentials: A Review of Research on Hedonic and Eudaimonic Well-Being. *Annual Review of Psychology*, *52*(1), 141-166. https://doi.org/10.1146/annurev.psych.52.1.141

Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *J Pers Soc Psychol*, *69*(4), 719-727. https://doi.org/10.1037/0022-3514.69.4.719

Chapter 3: Introduction to Psychometric Issues and Methods Used

Despite a clear need to measure adolescent mental health on a large scale by asking young people themselves, available self-report measures have been developed to poor standards (see Chapter 1). Often this means psychometric evidence is largely lacking (Bentley et al., 2019; Flake et al., 2017), suggesting work to gather this is needed. Developing new measures and refining constructs is a considerable task (Irwing & Hughes, 2018; Slaney, 2017). In addition, there can be reluctance to move away from canonical measures, since these are seen as valuable for making comparisons (Ford et al., 2020; Vostanis, 2006; Wolpert, 2020). This perception perhaps persists because of a lack of awareness of issues highlighted in Chapters 1 and 2. The comparison priority could also specifically represent flawed reasoning if, for instance, measures are interpreted and function differently in successive cohorts (Wicherts et al., 2004). In order to more robustly model the psychometric characteristics of measures known to be validated against a poor backdrop, and highlight potential issues to inform future work, the current thesis used a range of psychometric analyses. These were used to look for opportunities to use existing measures in more robust ways as well as highlight issues that should be addressed and inform future development. An overview is provided in this chapter to introduce relevant issues for the methods used in each of the papers, and how these techniques complement one another.

Secondary Data Analysis

Given the need to shed light on existing approaches, the current thesis drew on secondary analysis and review methodology. This meant large datasets could be drawn on and different measures considered (four general mental health measures were analyzed in the empirical studies). Given the number of parameters estimated in many psychometric analyses, large samples are needed (Epskamp, 2020b; Wang & Rhemtulla, 2021), again suggesting secondary analysis was an appropriate tool. The problems highlighted with the standards of development in adolescent mental health measurement (see Chapter 1) also mean it was important to consider multiple measures through multiple lenses. Secondary analysis was therefore ideal. This approach also strengthens generalizability, since though different measures often claim to measure the same construct, it is not always expected that findings will be uniform across different measures (Forbes et al., 2017; Rodebaugh et al., 2018).

While a key limitation of secondary analysis is often that data were not explicitly collected for the study in question (McCall & Appelbaum, 1991; Weston et al., 2019), this is less of a consideration in the current thesis: The focus here was on interrogating existing measures as they are routinely used in current research, and the fact multiple measures and samples could be drawn on again supported this aim. Another key issue for secondary analysis is potential overfitting (Weston et al., 2019). This could arise, for instance, through familiarity with the data leading to particular hypotheses being selected, using statistical tests for which the assumption is hypotheses have been selected a priori, or inappropriate testing and reporting of subgroup analyses (Weston et al., 2019). The empirical papers of the current thesis, which all used secondary data, were not preregistered. Doing so is often considered controversial (Weston et al., 2019) with existing frameworks argued to be poorly suited to secondary data. Indeed, recommendations to overcome inherent challenges have only recently been published, since the empirical papers of the current thesis were completed (Baldwin et al., 2022).

Nevertheless, several steps were taken to avoid overfitting and type I errors: First, in Papers 1 and 3 analytical steps were clearly reported and followed best practice based on simulation evidence where available. This included issues such as robust estimation accounting for the categorical and clustered nature of the data where possible and appropriate. Papers 1 and 3 considered construct and measure-level questions, meaning there were relatively limited options for how variables were treated, particularly when following simulation evidence (e.g., Li, 2016). These analytical considerations, e.g., estimator choices, were the primary researcher degrees of freedom. The transparent reporting of these based on prior literature means others can assess their suitability.

Second, since data were analyzed at the item level, extreme values were not considered and no data were excluded. Though there is no way to verify other preprocessing steps were not experimented with, this consistent approach across papers was adopted and reported in the interest of transparency. Similarly, since secondary data are typically recommended as an effective means to provide statistical power, this was not explicitly checked. Guidance also tends to focus on this benefit, rather than assessing potential oversensitivity (Kievit et al., 2022), which could have preprocessing implications. Cases were therefore not removed based on concerns about being overpowered. This is also consistent with how secondary analyses are typically conducted (e.g., Orben & Przybylski, 2019). Nevertheless, this potential risk was accommodated in the current thesis by not interpreting chi-square model fit where possible (see below), and considering the meaning of parameter estimates in context, including sensitivity analyses and theoretical issues. From a transparency and preprocessing perspective, consistently maintaining total samples can also be seen as a benefit.

Third, a multiverse approach, in which all justifiable approaches are analyzed, was adopted in Paper 2 since there were a larger number of analytical options, given the topic and methods. This allowed exploration of the sensitivity of results to different specifications (Steegen et al., 2016).

Fourth, while subgroup data were used (selected waves from a larger longitudinal study), this was based on pragmatic decisions such as which data were available at the time of analysis (for more detail on samples see also Chapter 4). For the dataset used in Paper 1, only a single wave was available. For Paper 2, three waves were needed to fit the panel network model (Epskamp, 2020b) so all that were available at the time were used. For Paper 3, though more waves were available, the first wave was selected to allow inclusion of the youngest recommended age for the SDQ, age 11 (the HeadStart design only included this age group in the first wave).

Fifth, relevant results and publications relating to other analyses of the same data are acknowledged in the papers where appropriate and available at the time of publication, and included in Appendix 1. This makes clear any prior knowledge of the data so that this can be considered alongside papers.

Finally, to aid transparency, clear data access instructions are published with the empirical papers, and synthetic data and code are published alongside Paper 2. The increased open materials and open science practices in general for Paper 2 (which was finished after Papers 1 and 3) reflect a combination of issues relating to the specific questions in each paper, but also increased general attention to these issues even in the time that I worked on my thesis (Nosek et al., 2022), as well as my increasing expertise in this area. For consistency, code for the analyses in Papers 1 and 3 is now also presented in Appendices 2 and 3 respectively.

Analyzing secondary data also meant I was impartial to the data collection procedures and measures in the sample. Though I was not involved in the collection of any of the data used in the thesis, I was involved in a similar project (Humphrey et al., in press) which informed my interest and understanding of the topics considered. For instance, though the rationale for the readability paper was rooted in existing literature (Patalay, Hayes, et al., 2018), visiting schools and administering similar measures for the Good Behaviour Game trial increased my awareness of this as a potential issue.

Psychometric Validity

Though definitions of validity in psychometrics have lacked consensus, Hughes (2018) has argued two questions are key:

> "1. Am I measuring what I want to measure?
>
> 2. Is my measure useful?" (p. 752).

These questions are directly relevant to problems identified thus far, namely concerns about development standards and possible confusion and overlap between constructs (Chapter 1).

As argued in Chapter 1, psychometric standards may have been especially low in adolescent mental health. This is likely due to a combination of factors, including a historical lack of faith in adolescents' views. While basic information such as internal consistency coefficients is typically available, more fundamental analysis considering the validity of items and constructs is typically lacking (Bentley et al., 2019). This is likely because measures' validity and reliability tend to be boiled down to a couple of basic heuristics, such that minimal information is reported and as such has been accepted as the norm (Flake & Fried, 2020; Flake et al., 2017). More insight into this issue, and the current state of psychometric evidence in adolescent general mental health, is also provided in Paper 4. Another issue is that there is an inherent tension between validity and reliability at the statistical level, such that though reliability is often preferenced in the literature, this can actually come at the cost of validity (Alexandrova & Haybron, 2016; Clifton, 2020).

There is therefore a clear need to consider validity more thoroughly. The following elements of psychometric validity are introduced below: Content validity relates to conceptualization and operationalization, while structural and external validity relate to empirical models (Flake et al., 2017; Loevinger, 1957). As argued in the previous chapters, both theoretical and empirical investigation is lacking. However, these must be balanced. Prioritizing empirical study and avoiding theory could mean important components are omitted (Alexandrova & Haybron, 2016), while a lack of empirical work has important implications for the interpretation of constructs and scores (McNeish & Wolf, 2020).

Content Validity

The most fundamental building block of a measure's psychometric properties is content validity, "the relevance, comprehensiveness, and comprehensibility of the [measure] for the construct, target population, and context of use of interest" (Terwee et al., 2018, pp. 1159-1160). Crucially then, measures and constructs should be developed in consultation with adolescents, for instance making sure items reflect relevant and comprehensible content (Deighton et al., 2013). Without this facet of validity, others are uncertain (Mokkink et al., 2018).

However, content validity evidence is resource-intensive to collect, involving qualitative work, literature searching, and cognitive interviews (Terwee et al., 2018). Given the focus on existing measures, content validity was considered in the following ways in the current thesis: In Paper 1, measures which had undergone some content validity testing during development (Deighton et al., 2013; Duncan et al., 2006) were used. This was important since the paper is largely focused at the construct level. Paper 3 considered age-appropriateness via readability and measurement invariance across age (whether items were responded to in similar ways by different age groups) to provide insight into the content validity of a self-report mental health measure. For Paper 2, items were selected partly based on readability findings from Paper 3 (the papers are not presented chronologically, see Chapter 4), and sensitivity to item operationalizations also provided adjunct insight into content validity. Finally, in Paper 4, content validity of available measures was rated according to established criteria (Terwee et al., 2007).

Structural Validity

Structural validity is concerned with the relationships between items and constructs, for instance via item-total correlations, factor analysis, or most recently, network psychometrics (Christensen et al., 2020; Flake et al., 2017; Hughes, 2018). Structural validity in the current thesis was investigated through the modeling of empirical relationships between symptom and wellbeing indicators and constructs, which had previously been analyzed in more limited ways (see Chapter 1 and Papers 1 and 2).

Though established procedures for considering structural validity have been recommended (e.g., Flake et al., 2017; Hughes, 2018), it has recently been highlighted that the statistical models for doing so should be interpreted cautiously. In fact, researchers are faced with a host of potential pitfalls. For instance, Rhemtulla et al. (2020) referred to a mismatch between theoretical constructs and statistical models as “construct invalidity”. This mismatch was further discussed by Fried (2020) who argued that the usually flexible or poorly defined theories associated with factor and network covariance models are not falsifiable by the statistical approaches typically deployed.

A more specific example relevant to the current thesis is that the presence of multiple correlated factors (based on model fit) should not be considered as evidence of multiple separate constructs without also considering the extent of dimensionality (Gignac & Kretzschmar, 2017). Considering the extent of dimensionality allows pragmatic insight beyond model fit, for instance on how highly correlated constructs or item clusters should be treated (Stochl et al., 2020). This is also important since structural analysis of indicators does not provide direct or objective evidence of how constructs are organized, with, for example, subjective decisions taken by researchers having a substantial bearing on results (Haeffel et al., 2021).

Factor and network models should therefore be considered alongside wider conceptual issues, and statistics beyond model fit. Not doing so could result in faulty coverage of the construct where items are included/removed based on entirely data-driven methods (Alexandrova & Haybron, 2016), or biased parameter estimates where the model is inappropriate (Neal & Neal, 2021; Rhemtulla et al., 2020).

Nevertheless, though covariance models cannot provide direct evidence of how systems are truly organized, they can provide insight that could further understanding (DeYoung et al., 2021; DeYoung & Krueger, 2020), and inform use of measures (e.g., Stochl et al., 2020). Given the need for exploratory work highlighted in Chapter 1 and the issues highlighted above, covariance models were considered. However, these were assessed in the context of wider issues including unidimensionality assessment, readability, item quality, item operationalization, and content.

External Validity

The term external validity has been used in psychometrics to refer to the stage in which a target measure is statistically compared to other variables, e.g., via correlation (Flake et al., 2017; Loevinger, 1957). However, similar to the inferential leap described from statistical models to theory (Fried, 2020), such correlational analyses of constructs with external variables provide only “circumstantial evidence” of construct validity (Borsboom et al., 2004, p. 1062). Standards for determining what constitutes a strong or weak enough relationship to determine, for instance, convergent or divergent validity, are also not clear and are therefore at the researcher's discretion (Mokkink et al., 2018). Together these considerations suggest external validity testing can be problematic. Nevertheless, the importance of other systems and psychological processes for the development of mental health (Bronfenbrenner, 2005; Rutter & Sroufe, 2000), and availability of additional variables in the datasets worked with, meant consideration of relationships to external variables was useful. Therefore, rather than correlational or receiver operator curve analyses of scores, which might be underpinned by poor measurement practices threatening conclusions (Flake & Fried, 2020), relationships to external variables were considered more critically at the indicator level via a multiverse framework (Paper 2).

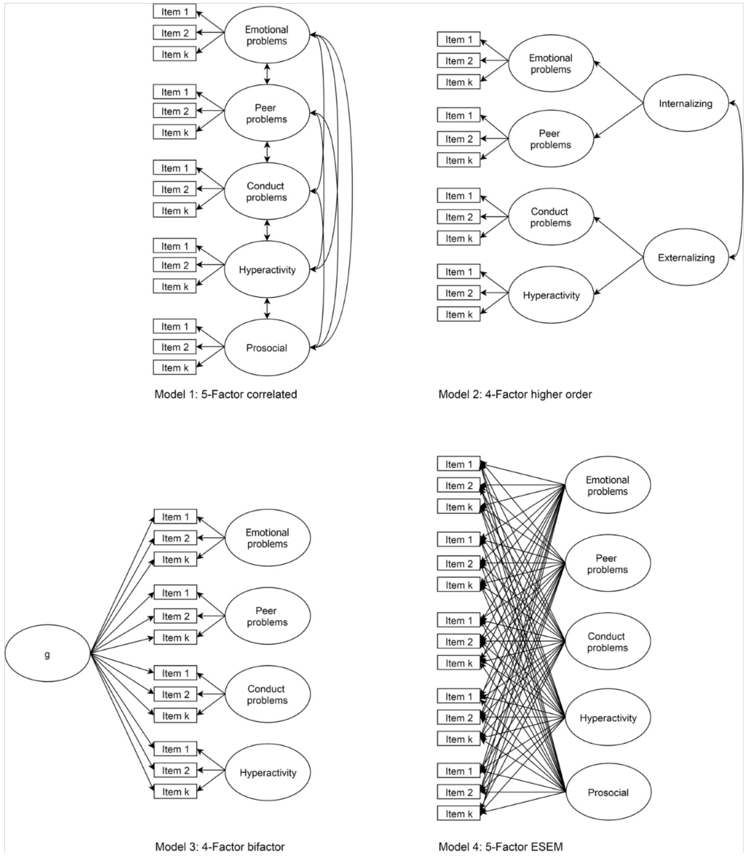
Empirical Psychometric Modeling

Several psychometric modeling techniques were used to explore the aspects of validity outlined above. Confirmatory factor analysis, cross-lagged panel network modeling, and exploratory structural equation modeling (ESEM) were used to consider construct-level relationships, dimensionality and measurement invariance across gender and socio-economic status (Paper 1), indicator-level relationships (Paper 2), and factor structure and measurement invariance across age (Paper 3). These methods are described in the papers of the thesis. Figure 3.1 reproduces figures from Papers 2 and 3 and provides a graphical illustration of the various modeling approaches used. Models 1-4 in the left panel are ordered with increasing parameterization, with latent factors represented by ovals which cause observed items (represented by rectangles). The right panel displays a temporal network model and is data-driven, with items (circles) causing one another in complex ways. However, a few overriding considerations that apply across methods and to their combined use are presented here.

Though network analysis has arisen in part as a reaction against some of the assumptions of latent variable models, and the two are often pitted against one another theoretically (Borsboom & Cramer, 2013), factor and network models are statistically closely related (Epskamp et al., 2017), and are subject to similar data/theory trade-offs (Fried, 2020). They can also provide similar insight. For instance, some centrality metrics, which provide insight into how frequently and strongly indicators covary with all others, have been shown to provide almost identical information to factor loadings (Hallquist et al., 2019). Factor and network methods can also both be used in an exploratory way when theory is unclear about the structure of indicators and constructs (Fried, 2020), as was the case here. Furthermore, I argue they can be used together to provide complementary insight. Considerations and background for the (joint) use of these methods are provided in the following sections.

Figure 3.1

Example Psychometric Models Used

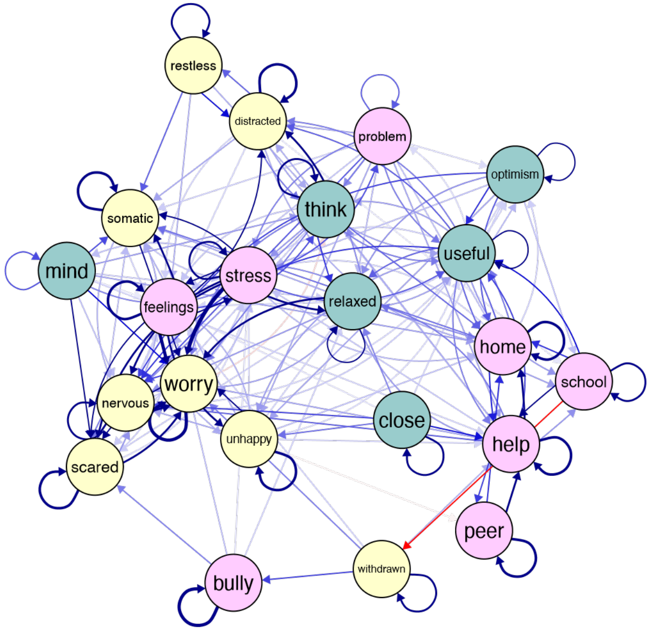


Model 1: 5-Factor correlated

Model 2: 4-Factor higher order

Model 3: 4-Factor bifactor

Model 4: 5-Factor ESEM



CFA and ESEM Models Estimated in Paper 3

Temporal Network from Paper 2

Factor Analysis

The common factor model parses the variance of observed item responses into shared variance (with other items) through a latent factor, and unique variance, a residual which encodes variance specific to the target item (Brown, 2015). CFA can be used to examine whether a set of a priori latent variables account for a specified pattern of shared variance among a set of observed indicators (Brown, 2015; DeYoung & Krueger, 2020). In clinical or personality psychology such latent variables could be depression or openness respectively. Indicators are typically items on questionnaires designed to measure the target construct. Item response theory models can be used in similar ways to model relationships between constructs and indicators, as well as consider dimensionality (Stochl et al., 2020). However, CFA was selected in the current thesis since this approach was often used in other similar work (e.g., Keyes, 2005; Patalay, Fonagy, et al., 2018), therefore allowing some comparison.

Some suggest that the strong assumptions of CFA mean it should not be used to summarize groups of items and that instead principal components analysis (PCA) is better suited to this end (Fried, 2020). However, PCA focuses on entirely data-driven dimensions which are not specified a priori. CFA, therefore, has two advantages over PCA in relation to the gaps addressed in the current thesis: First, it is likely positive and negative mental health constructs and indicators will cohere more within measures, and therefore domains, than between, in part due to wording and instrument effects (Clifton, 2020; Weijters & Baumgartner, 2012). It was therefore important to be able to specify a priori models and not be entirely driven by these potential data artifacts. Second, CFA can be used to assess essential unidimensionality via bifactor models and related indices (Reise et al., 2016; Rodriguez et al., 2016; Stochl et al., 2020). This was of interest in the current thesis since the similarity and dissociation between positive and negative mental health were somewhat unclear in prior literature (see Chapter 1 and Papers 1 and 2). The use of CFA in the current thesis was therefore pragmatic, since it facilitated consideration of a priori constructs, their relationship, and unidimensionality.

However, two stringent restrictions of CFA models are relevant to the current thesis: the assumption that items are locally independent after accounting for any shared variance via the latent factor; and the assumption that items cannot load on more than one factor. Both are substantively important. The first because items with similar wording and content are likely to be related beyond any underlying trait that causes both (Cramer et al., 2012). As discussed in Paper 1, residual covariances should therefore be considered. Detail is provided in Paper 1, but broadly I emphasize here, in line with a strong theme of the opening chapters, that this must be done with reference to theory. Entirely data-driven approaches to identifying error covariances merely for improving model fit are undesirable, and considering parameters sequentially leads to multiple testing and accuracy problems (Epskamp et al., 2017; Pan et al., 2017). Similarly, the issue of cross-loadings may be particularly important where constructs within a measure are not well defined. As discussed in Paper 3, this can be addressed via the use of ESEM which estimates factors but relaxes the cross-loading assumption (Asparouhov & Muthén, 2009; Marsh et al., 2014). Both these CFA restrictions needed explicit consideration in the current thesis since the development and conceptual problems highlighted in Chapters 1 and 2, suggest items and constructs may not have received sufficient attention.

Network Modeling

As mentioned above, network methods have emerged in part as a reaction against the restrictions of CFA (Cramer et al., 2012; Epskamp et al., 2017). They are data-driven, typically estimating partial correlations (Epskamp et al., 2018; Robinaugh et al., 2020), but have a broad underlying theory: symptoms and experiences in mental health likely influence one another in complex ways, and particular patterns of connectivity may give rise to disorder (Borsboom, 2017). A key idea is that symptoms cause one another. For instance, a lack of sleep might cause concentration problems which in turn might cause low mood. This is an alternative explanation for these symptoms' covariance to the diagnostic and dimensional approaches which hold that each of the symptoms is caused by a latent disease or dimensional process. While diagnoses and dimensional constructs (e.g., internalizing symptoms) are often used in research, likely at least in part as they afford methods that can be easily entered into models (e.g. sum-scores into regression or CFA in structural equation modeling), networks have the potential to offer an alternative. For instance, networks could be more consistent with clinical approaches such as formulation (von Klipstein et al., 2020). Nevertheless, network methods are at an early stage and such implications are yet to be fully worked out methodologically and theoretically (Fried, 2020; Rhemtulla et al., 2020; Robinaugh et al., 2020).

In addition, network models are subject to several other considerations. First, it is a fast-paced field in terms of empirical and methodological work (Robinaugh et al., 2020). For instance, cross-sectional network papers have proliferated, and have quickly come under criticism (Robinaugh et al., 2020). While some work suggests they can successfully approximate wider within-person effects which require longitudinal data (von Klipstein et al., 2021), other evidence suggests the opposite (Bos et al., 2017). Such cross-sectional analyses are therefore difficult to interpret. Since network theory posits that symptoms influence one another within rather than between individuals (Fried, 2020), this should most likely be analysed via longitudinal designs.[3] For this reason, I have not presented the cross-sectional network analysis I undertook during my PhD as part of the thesis. However, as recommended for transparent secondary analysis when the same dataset is used (Weston et al., 2019), I present it in Appendix 1. Paper 2 instead utilises the cross-lagged panel network model, which enabled consideration of average within-person processes via longitudinal modeling.

The specific panel network model, and related indices used in Paper 2, are introduced in some detail in the paper. However, in terms of the considerations for validity outlined above, the following background is useful. While methods were not available to estimate relationships between network measurement models as a whole and proximal variables via structural relationships (Rhemtulla et al., 2020), models such as the panel network allow consideration of such variables *within* the network. For instance, Isvoranu et al. (2020) explored how genetic markers interact with psychosis symptoms. In this sense, structural relationships can be explored between individual indicators and the target variable. This does not provide evidence for or against the network system, as it might be argued considering relationship between external variables and scores or latent variables could (Flake et al., 2017). However, as discussed above, such correlational evidence between scores and external variables is problematic with unclear standards and interpretations (Borsboom et al., 2004; Mokkink et al., 2018). Rather, including variables beyond the construct under study in a network may help clarify which indicators particularly

[3] While some argue that latent variable approaches are also inherently causal models, they can be considered tools for considering variance (DeYoung et al., 2021), and they do so in a much more constrained way than data-driven network models.

influence or are influenced by related phenomena. In line with the identified need in this field (see Chapter 1), such work is therefore exploratory.

Analytical Choices

A crucial element of designing psychometric analyses is selecting estimation procedures and criteria to judge models most likely to lead to accurate results and conclusions. Simulation literature is available to support these decisions but such papers are never entirely bespoke to a given study (McNeish & Wolf, 2021). An overview of how choices were made in the papers is therefore provided here.

Estimation

Whereas a blanket approach was taken for treatment of extreme values and power (see above), partly in the interest of transparency, this was not appropriate for selecting model estimators for a number of reasons, despite the fact similar ordinal data were considered throughout Papers 1-3. For instance, choice was much more limited for the panel network in Paper 2 since the software and method are much newer, and there was substantial missing data associated with the longitudinal design. Similarly, there were differing priorities between Papers 1 and 3 (which both used factor analysis) with the former explicitly seeking to handle error covariances and the latter focusing more on invariance testing through a less restrictive approach.

Broadly, where possible, the primary aim was to account for the ordinal or non-normal nature of the self-report item data used. Only the Outcome Rating Scale used in Paper 1 produced continuous responses, while the other measures ranged between three- and five-point Likert scales. Prior to conducting Paper 1, there was only limited evidence to inform estimator choices for mixed categorical and continuous data (Li, 2021). Therefore, as described in the paper, handling error covariances, the large sample size and low missingness suggested the weighted least squares means and variance adjusted (WLSMV) estimator should be used (Li, 2016; Muthén et al., 2015). Since the publication of this paper, new simulation evidence also suggests WLSMV is best suited to handling mixed continuous/categorical data (Li, 2021), further supporting the choice made.

While Paper 1 included measurement invariance testing, this was not the primary focus of the paper. On the other hand, for Paper 3, measurement invariance across age was the driving rationale for analysing factor models. Given this, despite the fact ordinal data were used, the robust maximum likelihood (MLR) estimator was selected. As described in the paper, this allowed use of more established means for model comparison via approximate fit differences, which was necessary due to the large sample size (see also discussion of chi-square tests in the section below; Sass et al., 2014). In light of this compromise, the main models were also considered using WLSMV for sensitivity.

Considerations for Papers 1 and 3 make clear that though no simulation is bespoke to a given empirical scenario, there is considerable literature on which to base decisions for factor models. In contrast, the available evidence for the panel network model of Paper 2 was limited to the paper that introduced it, and no estimator to treat data as ordinal was available (Epskamp, 2020a; Epskamp, 2020b). The full information maximum likelihood estimator was therefore chosen which was appropriate to handle the more substantial missing data, as well as skewed distributions (Muthén et al., 2015).

Model Fit

Covariance models can be statistically tested and compared using the chi-square test. This considers whether the collective differences between the actual and model-implied covariance matrices are significantly different from zero (Barrett, 2007). This has been argued to be problematic since it is extremely unlikely that *any* restrictive structural equation model would show this kind of exact fit (Steiger, 2007). Furthermore, since the result is a multiplier of sample size, models, including data-driven models such as networks, run in larger samples are more likely to show statistically significant differences and "fail" the test (Barrett, 2007). As described already, large sample sizes are needed for the models considered in the current thesis, and therefore chi-square results are typically reported but not interpreted as indicative that models should be rejected. This approach was adopted in Papers 2 and 3, and conservative cut-offs for alternative fit indices were used alongside other model considerations. This approach was also used to compare structures in Paper 1, but chi-square difference testing was used for measurement invariance testing, given that the WLSMV estimator was preferred over MLR (which would have allowed for established criteria in approximate fit difference testing; Sass et al., 2014).

A number of alternative approximate fit indices exist with established thresholds (Hu & Bentler, 1999). A comprehensive review of evidence for and against various fit indices is beyond the scope of this chapter. Nevertheless, a few considerations are highlighted as context for the papers. First, a key limitation is that no universal cut-offs can be determined since the behaviour of indices is affected by properties of models and data (McNeish & Wolf, 2021). This means that published thresholds will lead to over and under-rejection in certain cases (e.g., Xia & Yang, 2019). To address this, it is recommended that the chi-square result, degrees of freedom, sample size and some descriptive statistics should be reported so that results can be interpreted in light of issues that might affect fit (Markland, 2007). Second, different fit indices provide different information and these can therefore be used together to better understand the appropriateness of models (Miles & Shevlin, 2007). Disagreement between fit indices, i.e. if one meets a cut-off but another does not, also provides additional information and suggests that potential problems should be carefully evaluated (Crede & Harms, 2019; Lai & Green, 2016). Third, incremental fit indices overcome some of the issues associated with chi-square testing since these compare the model of interest with a null model, such that both are affected by sample size or reliability, cancelling this out (Miles & Shevlin, 2007).

Given the canonical status of the cut-offs suggested by Hu and Bentler (1999), and a lack of availability of bespoke approaches for the complex models estimated in the current thesis (bifactor, ESEM, network; McNeish & Wolf, 2021), these cut-offs were used. These cut-offs are also more conservative than others set out in the literature (Markland, 2007). Given the flexibility of ESEM (Paper 3) and the data-driven nature of the panel network model (Paper 2), two additional model selection indices that consider parsimony were also used for Papers 2 and 3, the Akaike information criterion (AIC) and Bayesian information criterion (BIC). These can be used for model comparison, and given the data-driven approach inherent to network models, may be particularly valuable here (Kan et al., 2019). In Paper 3, the Hu and Bentler (1999) canonical thresholds were used as a conservative benchmark to help adjudicate whether measurement invariance could proceed, i.e., to select the structure used to conduct measurement invariance analyses. The flexibility of the ESEM model was also explicitly acknowledged and though parameters provided insight into properties of the measure, the ESEM framework was primarily used to consider measurement invariance.

In addition, in each empirical paper, additional considerations beyond fit were taken into account so that fit was not used as a sole criterion: In Paper 1, the interpretability of models and additional indices to assess unidimensionality were evaluated; in Paper 2 the sensitivity of models to item operationalizations and estimation procedures was evaluated; and in Paper 3 item quality and readability were also used to draw conclusions about the measure and items.

Summary and Implications of Considerations for Using Empirical Psychometric Models

Psychometric models are powerful tools to understand the covariance structures of items. However, a multitude of choices are available to researchers, and the above sections make clear that these can be challenging. The following principles were adopted in the current thesis to navigate these issues: detailed, transparent reporting of modeling choices and fit; consideration of simulation evidence to inform this; evaluating sensitivity of models to analytical choices where appropriate (Papers 2 and 3); consideration of wider theoretical issues alongside estimating models (e.g., readability); evaluation of parameters (including derived indices) and parsimony as well as fit to ensure interpretability (see also issues for bifactor models described in Paper 1).

The first of these principles, transparency, is particularly important given the fast pace of the field, and use of secondary data analysis (Epskamp, 2019; Weston et al., 2019). While multiverse analysis is a powerful tool to address potential problems associated with novel methods and secondary data, this was done with relatively limited specifications as has been recommended (Del Giudice & Gangestad, 2021), and was not appropriate across the board. For instance, Paper 3 employed a WLSMV sensitivity model to check that the choice of the MLR estimator (preferred, given the availability of established invariance thresholds) was reasonable. However, given available simulation evidence and recommendations (Li, 2016; Muthén et al., 2015) there was no need to consider a wider range of estimators. In addition, while considering many reasonable approaches may help identify robust effects, multiverse results are also challenging to integrate for inference. For this reason, this approach was used sparingly in the current thesis, particularly given the numbers of parameters estimated in any given psychometric model.

The Need for Psychometric Approaches Beyond Statistical Models

The sections above demonstrate that factor and network modeling need to be treated carefully given problems such as the generalizability of fit cut-offs or estimation procedures, and the fact models do not provide direct evidence for theories. In light of this, several particular issues that might be missed by empirical models are identified here, and the methods used to address them in the papers of thesis are introduced.

First, the comprehensibility of a given measure and how it is interpreted, part of content validity should be considered. Consultation with stakeholders is important for measure development to define the construct, as well as check item wording. While this kind of direct insight into how adolescents interpreted items could not be achieved through the secondary datasets used in the current thesis, this was considered indirectly. Statistical modeling, measurement invariance analysis (Papers 1 and 3), afforded this by comparing how different groups responded, but also the multiverse design, since the sensitivity of effects to item wording was assessed (Paper 2). Where consultation with stakeholders is lacking in measures' development histories, checking readability also represents a quick exercise that provides some insight into comprehensibility. Insight into existing items could also be provided by checking these against standards for item development. Readability and item quality are complementary to one another and to empirical modeling, and were therefore applied together in Paper 3. For instance, if items have low comprehensibility, systematic noise could be introduced into any empirical model with the overall system or latent variables partly measuring, for instance, intelligence, rather than the construct of interest. The interpretation of models is therefore directly linked to comprehensibility.

A further issue, is the relative content of measures within and between constructs. Since it is known that findings often do not generalize well between measures (e.g., Rodebaugh et al., 2018), and there is likely conceptual overlap between general mental health domains (Alexandrova & Haybron, 2016), content analysis of indicators within and between constructs of general mental health is needed. This was conducted in Paper 4, following work in other domains which has suggested conceptualization is often inconsistent (e.g., Fried, 2017; Newson et al., 2020). Like comprehensibility, the consistency of conceptualization is key to interpreting empirical models, since substantive differences and similarities within and between constructs likely influence findings.

To advance the field of general adolescent mental health, issues beyond psychometric models must be considered, particularly as coarse and data-driven procedures have typically been employed to develop measures (see Chapter 1). In order to understand the validity of existing approaches, information about content and comprehensibility were therefore considered alongside empirical modeling.

Summary

This chapter has set out a tool kit for addressing some of the problems with measure and construct development in adolescent general mental health identified in Chapters 1 and 2. While new measures may be needed in the longer term, statistical models and theoretical approaches can be used together to shed light on the current state of play. This is needed to make clear the strengths and limitations on which to build.

Factor and network models can be used to provide insight into the covariance of indicators and constructs. However, inferences based on these models can be limited by the applicability of estimation procedures or cut-offs and appropriateness of a given model for a given theory or dataset. To safeguard against these issues, the current thesis prioritized transparency, used multiple statistical approaches and drew on available simulation evidence, as well as considering comprehensibility and content alongside statistical models.

References

Alexandrova, A., & Haybron, D. M. (2016). Is Construct Validation Valid? *Philosophy of Science*, *83*(5), 1098-1109. https://doi.org/10.1086/687941

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397-438. https://doi.org/10.1080/10705510903008204

Baldwin, J. R., Pingault, J.-B., Schoeler, T., Sallis, H. M., & Munafò, M. R. (2022). Protecting against researcher bias in secondary data analysis: challenges and potential solutions. *European Journal of Epidemiology*, *37*(1), 1-10. https://doi.org/10.1007/s10654-021-00839-0

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815-824. https://doi.org/https://doi.org/10.1016/j.paid.2006.09.018

Bentley, N., Hartley, S., & Bucci, S. (2019). Systematic Review of Self-Report Measures of General Mental Health and Wellbeing in Adolescent Mental Health. *Clinical Child and Family Psychology Review*, *22*(2), 225-252. https://doi.org/10.1007/s10567-018-00273-x

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5-13. https://doi.org/doi:10.1002/wps.20375

Borsboom, D., & Cramer, A. O. J. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91-121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061-1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bos, F. M., Snippe, E., de Vos, S., Hartmann, J. A., Simons, C. J. P., van der Krieke, L., . . . Wichers, M. (2017). Can We Jump from Cross-Sectional to Dynamic Interpretations of Networks Implications for the Network Perspective in Psychiatry. *Psychotherapy and Psychosomatics*, *86*(3), 175-177. https://doi.org/10.1159/000453583

Bronfenbrenner, U. (2005). *Making human beings human: Bioecological perspectives on human development*. Sage.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.

Christensen, A. P., Golino, H., & Silvia, P. J. (2020). A Psychometric Network Perspective on the Validity and Validation of Personality Trait Questionnaires. *European Journal of Personality*, *34*(6). https://doi.org/10.1002/per.2265

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, *25*(3), 259-270. https://doi.org/10.1037/met0000236

Cramer, A. O. J., Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., . . . Borsboom, D. (2012). Measurable Like Temperature or Mereological Like Flocking? On the Nature of Personality Traits. *European Journal of Personality*, *26*(4), 451-459. https://doi.org/doi:10.1002/per.1879

Crede, M., & Harms, P. (2019). Questionable research practices when using confirmatory factor analysis. *Journal of Managerial Psychology*, *34*(1), 18-30. https://doi.org/10.1108/JMP-06-2018-0272

Deighton, J., Tymms, P., Vostanis, P., Belsky, J., Fonagy, P., Brown, A., . . . Wolpert, M. (2013). The Development of a School-Based Measure of Child Mental Health. *Journal of Psychoeducational Assessment*, *31*(3), 247-257. https://doi.org/10.1177/0734282912465570

Del Giudice, M., & Gangestad, S. W. (2021). A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920954925. https://doi.org/10.1177/2515245920954925

DeYoung, C. G., Kotov, R., Krueger, R. F., Cicero, D. C., Conway, C. C., Eaton, N. R., . . . Wright, A. G. C. (2021). Answering Questions About the Hierarchical Taxonomy of Psychopathology (HiTOP): Analogies to Whales and Sharks Miss the Boat. *Clinical Psychological Science*, *0*(0), 21677026211049390. https://doi.org/10.1177/21677026211049390

DeYoung, C. G., & Krueger, R. F. (2020). To Wish Impossible Things: On the Ontological Status of Latent Variables and the Prospects for Theory in Psychology. *Psychological Inquiry*, *31*(4), 289-296. https://doi.org/10.1080/1047840X.2020.1853462

Duncan, B., Sparks, J., Miller, S., Bohanske, R., & Claud, D. (2006). Giving Youth a Voice: A Preliminary Study of the Reliability and Validity of a Brief Outcome Measure for Children, Adolescents, and Caretakers. *Journal of Brief Therapy*, *5*(2), 71-88.

Epskamp, S. (2019). Reproducibility and Replicability in a Fast-Paced Methodological World. *Advances in Methods and Practices in Psychological Science*, *2*(2), 145-155. https://doi.org/10.1177/2515245919847421

Epskamp, S. (2020a). *Package 'psychonetrics'*. https://cran.rproject.org/web/packages/psychonetrics/psychonetrics.pdf

Epskamp, S. (2020b). Psychometric network models from time-series and panel data. *Psychometrika*. https://doi.org/10.1007/s11336-020-09697-3

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195-212. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized Network Psychometrics: Combining Network and Latent Variable Models. *Psychometrika*, *82*(4), 904-927. https://doi.org/10.1007/s11336-017-9557-x

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456-465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378. https://doi.org/10.1177/1948550617693063

Forbes, M. K., Wright, A. G. C., Markon, K. E., & Krueger, R. F. (2017). Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology*, *126*(7), 969-988. https://doi.org/10.1037/abn0000276

Ford, T., Vizard, T., Sadler, K., McManus, S., Goodman, A., Merad, S., . . . Collinson, D. (2020). Data Resource Profile: Mental Health of Children and Young People (MHCYP) Surveys. *Int J Epidemiol*, *49*(2), 363-364g. https://doi.org/10.1093/ije/dyz259

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191-197. https://doi.org/https://doi.org/10.1016/j.jad.2016.10.019

Fried, E. I. (2020). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, *31*(4), 271-288. https://doi.org/10.1080/1047840X.2020.1853461

Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence*, *62*, 138-147. https://doi.org/https://doi.org/10.1016/j.intell.2017.04.001

Haeffel, G. J., Jeronimus, B. F., Kaiser, B. N., Weaver, L. J., Soyster, P. D., Fisher, A. J., . . . Lu, W. (2021). Folk Classification and Factor Rotations: Whales, Sharks, and the Problems With the Hierarchical Taxonomy of Psychopathology (HiTOP). *Clinical Psychological Science*, *0*(0), 21677026211002500. https://doi.org/10.1177/21677026211002500

Hallquist, M. N., Wright, A. G. C., & Molenaar, P. C. M. (2019). Problems with Centrality Measures in Psychopathology Symptom Networks: Why Network Psychometrics Cannot Escape Psychometric Theory. *Multivariate Behavioral Research*, 1-25. https://doi.org/10.1080/00273171.2019.1640103

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hughes, D. J. (2018). Psychometric Validity. In *The Wiley Handbook of Psychometric Testing* (pp. 751-779). https://doi.org/https://doi.org/10.1002/9781118489772.ch24

Humphrey, N., Hennessey, A., Troncoso, P., Panayiotou, M., Black, L., Petersen, K., . . . Lendrum, A. (in press). Examining the impact of the Good Behaviour Game on health- and education-related outcomes for children: a cluster RCT and cost-consequence analysis. *Public Health Research*.

Irwing, P., & Hughes, D. J. (2018). Test Development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing* (pp. 3-47). https://doi.org/10.1002/9781118489772.ch1

Isvoranu, A.-M., Guloksuz, S., Epskamp, S., van Os, J., & Borsboom, D. (2020). Toward incorporating genetic risk scores into symptom networks of psychosis. *Psychological Medicine*, *50*(4), 636-643. https://doi.org/10.1017/S003329171900045X

Kan, K.-J., van der Maas, H. L. J., & Levine, S. Z. (2019). Extending psychometric network analysis: Empirical evidence against g in favor of mutualism? *Intelligence*, *73*, 52-62. https://doi.org/https://doi.org/10.1016/j.intell.2018.12.004

Keyes, C. L. M. (2005). Mental Illness and/or Mental Health? Investigating Axioms of the Complete State Model of Health. *Journal of Consulting and Clinical Psychology*, *73*(3), 539-548. https://doi.org/10.1037/0022-006X.73.3.539

Kievit, R. A., McCormick, E. M., Fuhrmann, D., Deserno, M. K., & Orben, A. (2022). Using large, publicly available data sets to study adolescent development: opportunities and challenges. *Current Opinion in Psychology*, *44*, 303-308. https://doi.org/https://doi.org/10.1016/j.copsyc.2021.10.003

Lai, K., & Green, S. B. (2016). The Problem with Having Two Watches: Assessment of Fit When RMSEA and CFI Disagree. *Multivariate Behavioral Research*, *51*(2-3), 220-239. https://doi.org/10.1080/00273171.2015.1134306

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936-949. https://doi.org/10.3758/s13428-015-0619-7

Li, C.-H. (2021). Statistical estimation of structural equation models with a mixture of continuous and categorical observed variables. *Behavior Research Methods*, *53*(5), 2191-2213. https://doi.org/10.3758/s13428-021-01547-z

Loevinger, J. (1957). Objective Tests as Instruments of Psychological Theory. *Psychological Reports*, *3*(3), 635-694. https://doi.org/10.2466/pr0.1957.3.3.635

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modeling. *Personality and Individual Differences*, *42*(5), 851-858. https://doi.org/https://doi.org/10.1016/j.paid.2006.09.023

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory Structural Equation Modeling: An Integration of the Best Features of Exploratory and Confirmatory Factor Analysis. *AnnualReview of Clinical Psychology*, *10*(1), 85-110. https://doi.org/10.1146/annurev-clinpsy-032813153700

McCall, R. B., & Appelbaum, M. I. (1991). Some issues of conducting secondary analyses. *Developmental Psychology*, *27*(6), 911-917. https://doi.org/10.1037/0012-1649.27.6.911

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01398-0

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000425

Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences*, *42*(5), 869-874. https://doi.org/https://doi.org/10.1016/j.paid.2006.09.022

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonson, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2018). *COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) user manual Version 1.0*. COSMIN. https://cosmin.nl/wp-content/uploads/COSMINsyst-review-for-PROMs-manual_version-1_feb-2018.pdf

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2015). *Estimator choices with categorical outcomes*. Retrieved 10/12/2018 from http://www.statmodel.com/download/EstimatorChoices.pdf

Neal, Z. P., & Neal, J. W. (2021). Out of bounds? The boundary specification problem for centrality in psychological networks. *Psychological Methods*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000426

Newson, J. J., Hunter, D., & Thiagarajan, T. C. (2020). The Heterogeneity of Mental Health Assessment. *Frontiers in Psychiatry*, *11*(76). https://doi.org/10.3389/fpsyt.2020.00076

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., . . . Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(1), 719-748. https://doi.org/10.1146/annurev-psych-020821-114157

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*(2), 173-182. https://doi.org/10.1038/s41562-018-0506-1

Pan, J., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The Bayesian lasso. *Psychological Methods*, *22*(4), 687-704. https://doi.org/10.1037/met0000112

Patalay, P., Fonagy, P., Deighton, J., Belsky, J., Vostanis, P., & Wolpert, M. (2018). A general psychopathology factor in early adolescence. *British Journal of Psychiatry*, *207*(1), 15-22. https://doi.org/10.1192/bjp.bp.114.149591

Patalay, P., Hayes, D., & Wolpert, M. (2018). Assessing the readability of the self-reported Strengths and Difficulties Questionnaire. *BJPsych Open*, *4*(2), 55-57. https://doi.org/10.1192/bjo.2017.13

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the Bifactor Model a Better Model or Is It Just Better at Modeling Implausible Responses? Application of Iteratively Reweighted Least Squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research*, *51*(6), 818-838. https://doi.org/10.1080/00273171.2016.1243461

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30-45. https://doi.org/10.1037/met0000220

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R., & Borsboom, D. (2020). The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*, *50*(3), 353-366. https://doi.org/10.1017/S0033291719003404

Rodebaugh, T. L., Tonge, N. A., Piccirillo, M. L., Fried, E., Horenstein, A., Morrison, A. S., . . . Heimberg, R. G. (2018). Does centrality in a cross-sectional network suggest intervention targets for social anxiety disorder? *Journal of Consulting and Clinical Psychology*, *86*(10), 831-844. https://doi.org/10.1037/ccp0000336

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. *Journal of Personality Assessment*, *98*(3), 223-237. https://doi.org/10.1080/00223891.2015.1089249

Rutter, M., & Sroufe, L. A. (2000). Developmental psychopathology: Concepts and challenges. *Development and Psychopathology*, *12*(3), 265-296. https://doi.org/10.1017/S0954579400003023

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating Model Fit With Ordered Categorical Data Within a Measurement Invariance Framework: A Comparison of Estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167-180. https://doi.org/10.1080/10705511.2014.882658

Slaney, K. (2017). Construct Validity: Developments and Debates. In *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions* (pp. 83-109). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-38523-9_4

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702-712. https://doi.org/10.1177/1745691616658637

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*(5), 893-898. https://doi.org/https://doi.org/10.1016/j.paid.2006.09.017

Stochl, J., Fried, E. I., Fritz, J., Croudace, T. J., Russo, D. A., Knight, C., . . . Perez, J. (2020). On Dimensionality, Measurement Invariance, and Suitability of Sum Scores for the PHQ-9 and the GAD-7. *Assessment*, *0*(0), 1073191120976863. https://doi.org/10.1177/1073191120976863

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., . . . de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34-42. https://doi.org/https://doi.org/10.1016/j.jclinepi.2006.03.012

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., . . . Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research*, *27*(5), 1159-1170. https://doi.org/10.1007/s11136-018-1829-0

von Klipstein, L., Borsboom, D., & Arntz, A. (2021). The exploratory value of cross-sectional partial correlation networks: Predicting relationships between change trajectories in borderline personality disorder. *PLOS ONE*, *16*(7), e0254496. https://doi.org/10.1371/journal.pone.0254496

von Klipstein, L., Riese, H., van der Veen, D. C., Servaas, M. N., & Schoevers, R. A. (2020). Using person-specific networks in psychotherapy: challenges, limitations, and how we could use them anyway. *BMC Medicine*, *18*(1), 345. https://doi.org/10.1186/s12916-020-01818-0

Vostanis, P. (2006). Strengths and Difficulties Questionnaire: Research and clinical applications. *Current Opinion in Psychiatry*, *19*(4), 367-372. https://doi.org/10.1097/01.yco.0000228755.72366.05

Wang, Y. A., & Rhemtulla, M. (2021). Power Analysis for Parameter Estimation in Structural Equation Modeling: A Discussion and Tutorial. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920918253. https://doi.org/10.1177/2515245920918253

Weijters, B., & Baumgartner, H. (2012). Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research*, *49*(5), 737-747. https://doi.org/10.1509/jmr.11.0368

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for Increasing the Transparency of Analysis of Preexisting Data Sets. *Advances in Methods and Practices in Psychological Science*, *2*(3), 214-227. https://doi.org/10.1177/2515245919848684

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*(5), 509-537. https://doi.org/https://doi.org/10.1016/j.intell.2004.07.002

Wolpert, M. (2020). *Funders agree first common metrics for mental health science*. Retrieved 23/02/2022 from https://www.linkedin.com/pulse/funders-agree-first-common-metrics-mental-health-sciencewolpert

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409-428. https://doi.org/10.3758/s13428-018-1055-2

Chapter 4: Overview of Aims, Papers, and Data

Aims

Based on the issues laid out in Chapters 1 and 2, and methodological considerations in Chapter 3, the current thesis had the following broad aims: 1) consider the relationships between positive and negative mental health constructs and indicators using robust methods (Papers 1, 2 and 4); 2) consider correlates and development of positive and negative mental health (Paper 2); 3) examine measurement issues beyond empirical modeling (Papers 3 and 4); 4) based on 1-3, provide insight into more robust approaches to measuring general mental health (Papers 1-4). How these aims are operationalized in the papers and the choice of data for each is described below (more details about the data and research questions are provided in the papers themselves).

Relationship of the Papers to Each Other

Given the range of issues and possible approaches to meet the above aims described in the previous chapters, the current thesis adopted a journal format. How the results of the papers relate to one another is discussed in detail in the final chapter. Nevertheless, a brief overview of the links between papers is also provided here. The papers are not presented in chronological order, but are organized thematically. The first two papers both primarily explore structural issues when combining positive and negative approaches to adolescent general mental health. The final two papers provide insight into the wider issues of age appropriateness, conceptualization, and psychometric properties across the field.

Papers 1 and 2 respectively considered construct and indicator-level relationships for positive and negative aspects of general mental health in adolescence. Paper 2 also considered longitudinal relationships and inter/intra-personal correlates. Paper 1 influenced Paper 2, since the finding of a strong relationship between internalizing symptoms and wellbeing formed part of the rationale to focus on these domains in Paper 2.

Paper 3 provides detailed insight into age appropriateness for a widely used measure, the SDQ, while Paper 4 provides insight into the general psychometric and conceptual landscape via meta-review methodology. Paper 3 influenced Paper 2 with item choices in Paper 2 partly informed by the findings of

Paper 3. Paper 4 was conducted last and was influenced by each of the preceding papers. These made clear a need to evaluate and bring together conceptual and psychometric issues across general mental health in adolescence.

Paper 1

In the prior literature, the association between positive and negative mental health constructs had not been adequately modeled (see Chapter 1, and Paper 1 for more detail). This was addressed in Paper 1 via factor models (Aims 1 and 4). A large sample and more robustly developed measures were therefore important to handle the number of parameters, and ensure constructs were relatively well defined and operationalized. To meet these needs, the second phase of the HeadStart project was selected (Lereya et al., 2016). This allowed analysis of data from early adolescents with measures which had undergone qualitative work with young people to check their understanding of items and constructs (Deighton et al., 2013; Duncan et al., 2006).

Author Contribution. This paper was co-authored with Margarita Panayiotou (MP) and Neil Humphrey (NH). I came up with the initial idea for the paper and this was refined through discussion with MP and NH. I conducted all statistical analyses and drafted the paper. MP and NH commented on drafts.

Links to Other Papers. The strong relationship found between internalizing symptoms and wellbeing suggested a focus on these domains together, and the general internalizing distress factor also suggested a network approach might be justified (Paper 2).

Paper 2

Paper 2 explored the relationships between mental health indicators and inter/intra-personal correlates over time (Aims 1, 2, and 4). A large dataset with both positive and negative mental health indicators, as well as relevant malleable correlates at each time point was needed, with at least three relatively close time points (Epskamp, 2020). The longitudinal sample from the main phase of the HeadStart project (Deighton et al., 2019) was selected since it met these criteria.

Author Contribution. This paper was co-authored with MP and NH. I came up with the initial idea for the paper and this was refined through discussion with MP and NH. I conducted all statistical analyses and drafted the paper. I made initial suggestions for the multiverse conditions and these were agreed through discussion. MP and NH commented on drafts.

Links to Other Papers. This paper concentrated on internalizing symptoms and wellbeing via a network approach as suggested by the findings of Paper 1. The similarity of these constructs was again highlighted by similar complex relationships to correlates and centrality. In addition, indicator operationalization affected conclusions. Together, these issues suggested a need for insight into the conceptual and psychometric landscape of positive and negative mental health measures (Papers 3 and 4).

Paper 3

The Strengths and Difficulties Questionnaire is an extremely widely used measure (see Paper 3) which was developed without consultation with young people (Goodman, 1997; Goodman et al., 1998). In addition, evidence of readability issues had been identified at the subscale level (Patalay et al., 2018), and the measure is often used to describe age trends (see Paper 3). However, item-level readability, item quality, and any possible effects on different ages responding had not been considered (Aims 3 and 4). A large dataset which provided item responses across different age groups was therefore needed. The first year of the main, third phase of the HeadStart project met these criteria and was therefore used.

Author Contribution. This paper was co-authored with Rosie Mansfield (RM) and MP. I came up with the initial idea for the paper and this was refined through discussion with RM and MP. RM conducted the readability analysis. I conducted all statistical analyses and drafted the paper. RM and MP commented on drafts.

Links to Other Papers. Papers 1 and 2 made clear the need for accurate modeling of positive and negative mental health measures, and exploration of issues beyond this. This paper therefore employed robust modeling to consider structural issues and assessed readability and item quality. The paper demonstrated that omission of appropriate item development practices can have marked implications for the quality of a measure. This therefore again highlighted the need for a study providing a wide-ranging review of conceptual and empirical measurement issues (Paper 4).

Paper 4

Working on Papers 1-3 made clear that conceptualization and psychometric properties across measures in general mental health needed to be reviewed (Aims 1 and 4). Given the existence of relevant systematic reviews of measures (e.g., Deighton et al., 2014), but lack of robust psychometric/content analysis, a meta-review was conducted.

Author Contribution. This paper was co-authored with MP and NH. I came up with the initial idea for the paper and this was refined through discussion with MP and NH. I drafted the protocol, which MP and NH helped revise. I conducted the search and then MP and I both screened a 20% random subset of titles/abstracts in a pilot stage. Based on this, I screened the remaining records. MP and I both screened 100% of the full texts. Both title/abstract and full-text screening stages were supervised by NH. The content coding strategy was developed through discussion of a subset of indicators by all authors. I then coded all indicators which were checked for agreement by MP and also discussed with NH. I extracted psychometric properties and conducted COSMIN ratings in discussion with the other authors. I conducted the statistical analysis and drafted the paper. MP and NH commented on drafts.

Links to Other Papers. Each of the preceding papers highlighted measurement issues in specific measures. These included conceptual and empirical similarities not often accounted for (Papers 1 and 2); variation in results between different measure/item operationalizations (Paper 2); and structural and age-appropriateness issues in the SDQ (Paper 3). This paper, therefore, reviewed measurement issues in positive and negative mental health, mapping the content of items, measures, and domains and rating psychometric properties.

Why HeadStart and Not Other Secondary Data

While the current thesis aimed to capitalize on secondary data to draw on multiple samples and measures, all empirical papers used data from the HeadStart project. This was partly pragmatic, as data were available through working in the Manchester Institute of Education, and secondary studies such as those presented here were encouraged.

Nevertheless, other publicly and departmentally available datasets were considered but rejected since they were less suitable for the research questions of the current thesis. For instance, The Good Behaviour Game project, through which my PhD was funded, did not contain self-report mental ill health data (teacher-report was used instead), and focused mostly on children younger than 10. The Millennium Cohort Study was considered but this did not have self-report mental ill health data before age 14 (released at the start of my PhD; Patalay & Fitzsimons, 2017), or self-report SDQ responses. Positive and negative mental health were also measured differently at different time points, limiting the appropriateness for longitudinal modeling of the type conducted in Paper 2, and the positive measure focused on life satisfaction rather than the more comprehensive approach available in the main HeadStart waves (see Paper 3). Similarly, Understanding Society focuses on life satisfaction and measures fewer inter/intra-personal correlates than HeadStart (Jäckle et al., 2017).

Since conducting the analyses of the current thesis, Kievit et al. (2022) have created a resource on developmental adolescent datasets. Checking this against the following criteria confirmed other datasets were less suited to the current research than those used: freely available, non-categorical mental ill health measures, non-clinical samples, and self-report positive and negative mental health measures. In addition, though data all came from a single project, there was relatively little overlap with several samples and measures used (see Table 4.1).

Table 4.1

Overview of Data Used in the Empirical Papers

	Sample	Mental health Measures	Overlapping	*N*
Paper 1	HS pilot, age 10-11	M&MS CORS	No	1,982
Paper 2	HS age 11-12 baseline, T2, T3	SDQ SWEMWBS	Yes: baseline age 11-12 SDQ	15,843
Paper 3	HS baseline, age 11-12/13-14	SDQ	Yes: baseline age 11-12 SDQ	30,290

Note. HS = HeadStart; T2 = time two; T3 = time 3; M&MS = Me and My School; CORS = Child Outcome Rating Scale; SDQ = Strengths and Difficulties Questionnaire; SWEMWBS = Short Warwick Edinburgh Mental Well-Being Scale.

Summary

The papers of the thesis aimed to provide initial insight into robust approaches to measuring general mental health in adolescence. New methods and critical approaches were applied to improve understanding of psychometric and conceptual issues in four papers.

References

Deighton, J., Croudace, T., Fonagy, P., Brown, J., Patalay, P., & Wolpert, M. (2014). Measuring mental health and wellbeing outcomes for children and adolescents to inform practice and policy: a review of child self-report measures. *Child and Adolescent Psychiatry and Mental Health*, *8*(1), 14. https://doi.org/10.1186/1753-2000-8-14

Deighton, J., Lereya, S. T., Casey, P., Patalay, P., Humphrey, N., & Wolpert, M. (2019). Prevalence of mental health problems in schools: poverty and other risk factors among 28 000 adolescents in England. *The British Journal of Psychiatry*, 1-3. https://doi.org/10.1192/bjp.2019.19

Deighton, J., Tymms, P., Vostanis, P., Belsky, J., Fonagy, P., Brown, A., . . . Wolpert, M. (2013). The Development of a School-Based Measure of Child Mental Health. *Journal of Psychoeducational Assessment*, *31*(3), 247-257. https://doi.org/10.1177/0734282912465570

Duncan, B., Sparks, J., Miller, S., Bohanske, R., & Claud, D. (2006). Giving Youth a Voice: A Preliminary Study of the Reliability and Validity of a Brief Outcome Measure for Children, Adolescents, and Caretakers. *Journal of Brief Therapy*, *5*(2), 71-88.

Epskamp, S. (2020). Psychometric network models from time-series and panel data. *Psychometrika*. https://doi.org/10.1007/s11336-020-09697-3

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

Goodman, R., Meltzer, H., & Bailey, V. (1998). The strengths and difficulties questionnaire: A pilot study on the validity of the self-report version. *European Child & Adolescent Psychiatry*, *7*(3), 125-130. https://doi.org/10.1007/s007870050057

Jäckle, A., Gaia, A., Al Baghal, T., Burton, J., & Lynn, P. (2017). *Understanding Society The UK Household Longitudinal Study Innovation Panel, Waves 1-9, User Manual*. https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/innovationpanel/user-guides/6849_ip_waves1-9_user_manual_June_2017.pdf

Kievit, R. A., McCormick, E. M., Fuhrmann, D., Deserno, M. K., & Orben, A. (2022). Using large, publicly available data sets to study adolescent development: opportunities and challenges. *Current Opinion in Psychology*, *44*, 303-308. https://doi.org/https://doi.org/10.1016/j.copsyc.2021.10.003

Lereya, S. T., Humphrey, N., Patalay, P., Wolpert, M., Böhnke, J. R., Macdougall, A., & Deighton, J. (2016). The student resilience survey: psychometric validation and associations with mental health. *Child and Adolescent Psychiatry and Mental Health*, *10*(1), 44. https://doi.org/10.1186/s13034-016-0132-5

Patalay, P., & Fitzsimons, E. (2017). Mental ill-health among children of the new century: trends across childhood with a focus on age 14. September 2017. In. London: Centre for Longitudinal Studies.

Patalay, P., Hayes, D., & Wolpert, M. (2018). Assessing the readability of the self-reported Strengths and Difficulties Questionnaire. *BJPsych Open*, *4*(2), 55-57. https://doi.org/10.1192/bjo.2017.13

Paper 1: The Dimensionality and Latent Structure of Mental Health Difficulties and Wellbeing in Early Adolescence

Status: published, gold open access

Black, L., Panayiotou, M., & Humphrey, N. (2019). The dimensionality and latent structure of mental health difficulties and wellbeing in early adolescence. PLOS ONE, 14(2), e0213018. doi:10.1371/journal.pone.0213018

Supplementary material can be found in Appendix 2.

RESEARCH ARTICLE

The dimensionality and latent structure of mental health difficulties and wellbeing in early adolescence

Louise Black*, Margarita Panayiotou, Neil Humphrey

Manchester Institute of Education, University of Manchester, Manchester, United Kingdom

* louise.black@manchester.ac.uk

Abstract

Research with adults and older adolescents suggests a general factor may underlie both mental health difficulties and wellbeing. However, the classical bifactor model commonly used to demonstrate this general trait has recently been criticised when a unidimensional structure is not supported. Furthermore, research is lacking in this area with children and early adolescents. We present confirmatory factor analysis models to explore the structure of psychopathology and wellbeing in early adolescents, using secondary data from a large U.K. sample ($N = 1982$). A simple correlated factors structure fitted the data well and revealed that wellbeing was just as related to internalising as this was to externalising symptoms. The classical bifactor solution also fitted the data well but was rejected as the general factor explained only 55% of the total common variance. *S*-1 models were therefore used to explore general covariance in a more robust way, and revealed that a general internalising distress factor could play an important role in all item responses. Gender and income differences in mental health were also explored through invariance testing and correlations. Our findings demonstrate the importance of considering mental health difficulties and wellbeing items together, and suggestions are made for how their correspondence could be controlled for.

Citation: Black L, Panayiotou M, Humphrey N (2019) The dimensionality and latent structure of mental health difficulties and wellbeing in early adolescence. PLoS ONE 14(2): e0213018. https://doi.org/10.1371/journal.pone.0213018

Editor: Kenji Hashimoto, Chiba Daigaku, JAPAN

Received: July 9, 2018

Accepted: February 13, 2019

Published: February 26, 2019

Data Availability Statement: This study utilises secondary analysis of survey data collected for the pilot phase of the HeadStart (HS) learning programme, as well as demographic data extracted from the National Pupil Database (NPD). The authors are unable to make the NPD data available as they are restricted by their owners (the UK government). Demographic data (i.e. gender and free school meal eligibility) were obtained from the NPD (request numbers DR150407.01 and DR150818.01). More information about the process of requesting NPD data is available at the following URL: https://www.gov.uk/guidance/how-to-access-department-for-education-dfe-data-

Introduction

Both mental ill health and positive wellbeing in young people are associated with outcomes such as academic attainment and social functioning [1–5], as well as demographic and environmental correlates [6–14]. The majority of mental health problems have first onset in adolescence [15], and can result in significant disability [6, 8, 9]. Furthermore, it is widely agreed that adolescence, ranging from ages 10–24, is critical to functioning in later life [16–18], while recent evidence suggests young people's mental health may be deteriorating [6, 12].

Despite this clear need to understand the form of mental health, particularly in young people, its conceptualisation and measurement have been inconsistent. A historic focus on disorder remains the basis for measurement [19], even though the absence of disorder symptoms consistently fails to fully explain wellbeing in young people [1–5, 7, 13, 14]. The limitations of

extracts#apply-for-access-to-dfe-data. The HS survey data on mental health and wellbeing belongs to the Evidence Based Practice Unit (a collaboration between UCL and the Anna Freud National Centre for Children and Families, AFNCCF), who led the HS pilot evaluation. The authors accessed this survey data via membership in a consortium involved with the HS pilot evaluation. As collaborators on the main HS evaluation, the authors were granted secure remote access to this data by the principal investigator of the main HS evaluation, Dr Jessica Deighton. HS pilot data cannot be made publicly available, since consent was not obtained from participants for the public sharing of their survey responses. However, an anonymised version of the survey dataset used in the present paper is available on request from Dr. Jessica Deighton (Jessica.DeightonPhD@annafreud.org) or Dr. Tanya Lereya (Tanya.lereya@annafreud.org) under the following terms: 1. Schedule and arrange for site visit to AFNCCF to analyse data (password to user account supplied). 2. Analysis to be worked on in situ. 3. Results (but not data) taken away. In the event that either of these individual leaves the AFNCCF, updated contact information for new guardians of the data will be provided to PLOS.

Funding: The data used in this study was collected as part of HeadStart learning programme and supported by funding from the National Lottery Community Fund, grant R118420. The funders did not seek to influence any aspect of the secondary analysis reported in this study. The content is solely the responsibility of the authors and it does not reflect the views of the National Lottery Community Fund. Louise Black is funded by the National Institute for Health Research.

Competing interests: The authors have declared that no competing interests exist.

categorical diagnoses are also becoming increasingly clear, with criticisms focussing predominantly on stigmatisation via poorly evidenced medical models [19], and a lack of validity for discrete disorders [20, 21]. For instance, hyperactivity disorders have been criticised as pathologising typical and expected behaviour in children and adolescents, particularly boys [22], and studies have repeatedly failed to discern groups experiencing one externalising disorder without other comorbid problems [23–25]. Symptom-level and hierarchical approaches, on the other hand, are emerging as useful ways to understand structure, risk and comorbidity in mental health difficulties. Such approaches have demonstrated consistent covariance between symptoms, cutting across traditional disorder taxonomies [20, 26–32]. In fact, not only is there strong evidence of general covariance between symptoms of mental health, longitudinal research (from birth to midlife) suggests that experiencing symptoms of mental disorder is the norm, with only a small minority remaining completely symptom-free over time [33]. This supports the current shift in understanding, in which taxonomic approaches to mental disease classification are being rejected. Continuous dimensional frameworks are instead being adopted and encouraged, to reflect evidence that mental health symptoms seem to be extreme and distressing variations in typical processes rather than indicative of categorical diagnoses [34].

While *dual-factor* approaches have sought to gain a more comprehensive view of child and adolescent mental health by capitalising on the benefits of wellbeing measures [2], they too have typically resorted to simplistic categorical approaches. Though a moderate relationship between psychopathology and wellbeing has been consistently demonstrated [35–38], a focus has emerged which has emphasised their dissociation, forcing participants into one of four categories [1–5, 13, 14]. At either extreme, these are content and free of symptoms (*flourishing*), and dissatisfied and suffering symptoms (*languishing*). Also included, however, are the more surprising groups of individuals who are symptom-free and dissatisfied, and satisfied but symptomatic. This approach has demonstrated the important finding that absence of symptoms is not synonymous with the presence of wellbeing. However, it distracts from the known association between the two constructs, and finding that the majority of participants are straightforwardly either flourishing or languishing [1–5, 13, 14]. Nevertheless, wellbeing approaches do not appear to suffer from the outdated biases outlined above, and in young people there is also strong correspondence between different instruments and wellbeing subtypes, suggesting strong construct validity [10]. Given the association of mental health difficulties and wellbeing, the need for continuous approaches to mental health, and the relative strengths of wellbeing measures, there is therefore an opportunity to consider these outcomes together as part of a comprehensive structure.

Despite this, robust methods interrogating the measurement structure of wellbeing and mental health difficulties in early adolescence have yet to be employed, despite the existence of theoretical frameworks such as *complete mental health*, the *two-continua* approach, or the *dual-factor model* [2, 38, 39]. The current study addresses this major gap, building on research with adults and older adolescents [38, 40, 41].

Mental health difficulties and wellbeing

Wellbeing is typically considered to comprise positive (cognitive) evaluations of life, positive affect and the absence of negative affect [42]. These three aspects are typically considered to form hedonic wellbeing, while eudaimonic wellbeing captures aspects beyond pleasure, reflecting how well a person feels they align with their own values and ideals [43]. In young people, these different approaches to wellbeing have been shown to be highly related [10].

The present analysis draws on instruments designed for general population screening and will therefore focus on internalising and externalising symptoms. Though this means not all disorders and symptom-types are covered, this approach builds on previous research [7], provides insight into the two most common forms of mental health difficulties in childhood [8, 9], and is supported by evidence that broad internalising and externalising spectra can explain covariance across disorders [26].

Internalising is typically considered to include depressive and anxious type disorders and is therefore concerned with somatic, worry and sadness symptoms [26, 44]. There is, therefore, some conceptual crossover between this aspect of mental health difficulties and wellbeing, given that they are each is concerned with happiness or unhappiness. This can be seen in measures such as the General Health Questionnaire 12 (GHQ-12), which is sometimes considered to be a symptom measure, and sometimes a wellbeing instrument capturing negative affect [40, 45].

In children, externalising symptoms and disorders typically include conduct and attentional problems [46, 47]. Given the controversy surrounding attentional problems mentioned above, the current study focuses particularly on conduct problems. Though externalising symptoms often share comorbidity with internal distress symptoms, when considered alone these are behavioural and related to disinhibition [44].

Gender differences in child and adolescent mental health

The prevalence of disorders between genders is complex in each developmental period. Between ages 6 and 11 boys are up to twice as likely to suffer from severe mental health difficulties, but levels of internalising symptoms are similar [7, 8, 48]. However, between 11 and 14, girls are substantially more likely to suffer from internalising problems [6, 49]. Bifactor modelling has also yielded inconsistent results: While some research has suggested a general mental health factor was not associated with gender in early adolescence [28], a study with slightly older participants suggested it was [41]. The expression of mental health is therefore linked to gender in a complex way at the beginning of adolescence (around age 11), and warrants further investigation.

Wellbeing also shows consistent complex differences for gender, varying significantly by domain [10, 11]. Typically, girls show higher satisfaction with school and social relationships, while boys are happier with their appearance [11, 12]. Overall, wellbeing is higher for boys in some countries and for girls in others [11]. In the U.K., child and adolescent boys were shown to have higher overall happiness [12]. From a unidimensional perspective, this is incongruent with the finding in the same country that boys are at greater risk of mental health difficulties [48]. However, it perhaps echoes the finding that U.K. adolescent girls are at particular risk of depression [6, 49]. The complexity of gender relationships with mental health difficulties and wellbeing challenges assumptions of unipolarity, and suggests empirical evidence of their structure is needed.

Family income differences in child and adolescent mental health

Though country-level economic factors show no or very little association with children and adolescents' wellbeing or mental health difficulties, household-level income is significantly associated with these outcomes [6, 10, 11, 48, 50]. While patterns for income are more straightforward than for gender, with children from poorer backgrounds reporting greater mental health difficulties and lower wellbeing, the extent to which income explains each outcome is quite different. Family income consistently more strongly predicts variability in mental health difficulties than wellbeing [6, 10, 11, 48, 50]. The existence of this relationship for both outcomes in varying strength, suggests their composite structure may provide insight into the role of income for mental health.

Problems with the existing dual-factor approach

When mental health difficulties and wellbeing are analysed independently (i.e. any covariance is not accounted for), they do appear to be somewhat distinct. For instance, longitudinal research suggests that, even among the minority who never experience mental disorder, over 20% have been found to report low life satisfaction [33]. Similarly, the two constructs have been found to have a discrete set of correlates, as well as some shared predictors in early adolescence [7]. It remains unclear, however, to what extent items for each construct overlap and tap similar dimensions. For instance, while Patalay et al. [7] aggregated internalising and externalising symptoms (likely only moderately correlated; see [47]), and then found the corresponding coefficient between mental health difficulties and wellbeing to be only -.20, Kinderman et al. [51] treated wellbeing and internalising psychopathology as related latent factors, and these were correlated at -.82. The conceptual overlap between internalising and wellbeing alluded to above may explain this discrepancy between correlations since though both referred to outcomes as mental ill health, Kinderman et al. [51] included only depression and anxiety.

Given that mental health difficulties and wellbeing are known to be correlated, [37, 38], it seems illogical not to control for this association. Furthermore, since results are likely biased, already suggested by Patalay and Fitzsimons'[7] surprisingly low correlation between the two constructs and dimensionality is assumed rather than tested, conclusions based on analyses ignoring the association of mental health difficulties and wellbeing should be treated with caution.

Problems with existing approaches to modelling mental health

The definitions above make clear that mental health difficulties represent a broad range of symptoms, some of which intuitively relate to wellbeing, and that these constructs show complex relationships with gender and income. Complex measurement models are already common in mental health research since high rates of comorbidity and correlations between items have led researchers to model symptoms or disorders together through bifactor structures, termed psychopathology or *p*-factor models [27]. These models have been used to argue for a general transdiagnostic factor and two studies have extended these to include wellbeing [40, 41]. Despite appropriately controlling for wellbeing, these studies have focused on older samples and age generalisability cannot be assumed [6, 10, 48]. These studies also have theoretical and methodological problems leaving many questions unanswered. For instance, the study by Böhnke et al. [40] was restricted since the measure used for mental health difficulties (the GHQ-12) has been argued by some to mainly capture negative affect [45]. Therefore the finding by Böhnke et al. [40] of a strong general factor explained almost entirely by GHQ-12 indicators is arguably unsurprising, since this measure could be expected to strongly mirror wellbeing instruments [10, 45].

While Böhnke et al. [40] studied adults in the general population, St Clair et al. [41] aimed to understand the structure of mental health in a sample of older adolescents and young adults. While symptom measures were included, these tended to be old, based on categorical diagnoses, or poorly validated [52–55], and self-esteem was also included as a measure of positive mental health with no clear theoretical justification. This is therefore at odds with contemporary spectra approaches [26], and may explain why an arguably uninterpretable result emerged: The best fitting model was a bifactor solution, but items did not always load on both general and specific factors, some loadings were low and even reversed on specific factors, and crossloadings seemed to be allowed, such that wellbeing and self-esteem items were allowed to load on a shared positive factor as well as two separate specific factors. Eid et al. [56] point out that such problematic solutions can arise where bifactor models are misapplied, while the questionable choice of measures, unsupported by theory is likely to have contributed to the results outlined above. There is, therefore, a clear need to study the complex structure of mental health in adolescents using more appropriate measures.

Beyond these specific problems with dual-factor bifactor studies, there has recently been a great deal of criticism of bifactor modelling more generally, which the current study aims to address. Firstly, where there are correlations between all indicators, as is the case in mental health models, a general factor which accounts for this covariance will always occur, even where this pattern of covariance arises for another reason, such as network structures, where one symptom leads to another [57]. Secondly, bifactor structures are highly parameterised and tend to overfit the data such that sample and measure complexity (e.g. cross loadings and correlated residuals) can be absorbed by the general factor, making the bifactor structure apparently better fitting even when this is not the case [58]. Thirdly, though evaluating competing models is important to avoid selecting a model based on close fit alone, when others may be viable or better, model comparison between correlated factors, second-order and bifactor solutions as is typically conducted could lead to false conclusions [57–59]. While these structures have substantially different interpretations, they are mathematically very close and sometimes even equivalent (depending on the number of factors). As a result, differences may not be attributable to superior structure, but instead be an artefact of the sample, unmodeled complexity or an alternative explanation for covariance such as *mutualism* in which problems co-occur [57–59]. Relative fit of such models must therefore be interpreted with caution.

Recent criticisms have also proposed that the classical bifactor model (see Fig 1B) is not psychometrically well defined, since a single source of variability (the participants) is used to define a dual decomposition of a single score into two random variables, which ought to each have a distinct source of randomness [56]. This means that latent general and specific factors are unrelated while simultaneously being a function of the true score of the same indicators. Where these specific factors have substantial variance and salient loadings, these are therefore uninterpretable since they represent constructs that are wholly orthogonal to each other and the general factor, while this general factor simultaneously represents shared covariance [56, 60]. If we consider the general factor to represent liability for all symptoms, the residual specific factors must represent something wholly unrelated to the symptoms captured by the general factor [60]. On the other hand, if we consider a specific internalising factor to represent specific depressive, somatic and anxious symptomology, we must assume that the general factor does not include these in the same way. Given that both general and specific factors are generated from the same responses to the same item set, it is impossible to substantively distinguish these orthogonal true score variables as the constraints of the bifactor model require [56].

In order to estimate a meaningful general factor that captures the covariance of all items, one specific factor can be removed [56]. This allows the general factor to become a function of the true score of the items with no specific factor, so that it can become well defined psychometrically as a random variable. The general factor in this model, known as *S*-1 (see Fig 1C), however, has a slightly different interpretation. For instance, if the specific wellbeing factor is removed ($S\text{-}1_{\text{wellbeing}}$), the general factor represents general wellbeing accounting for the covariance of this construct with internalising and externalising items. The specific internalising and externalising factors, on the other hand, would represent the residual variance not explained in these items by the general wellbeing domain. We argue that this model should be

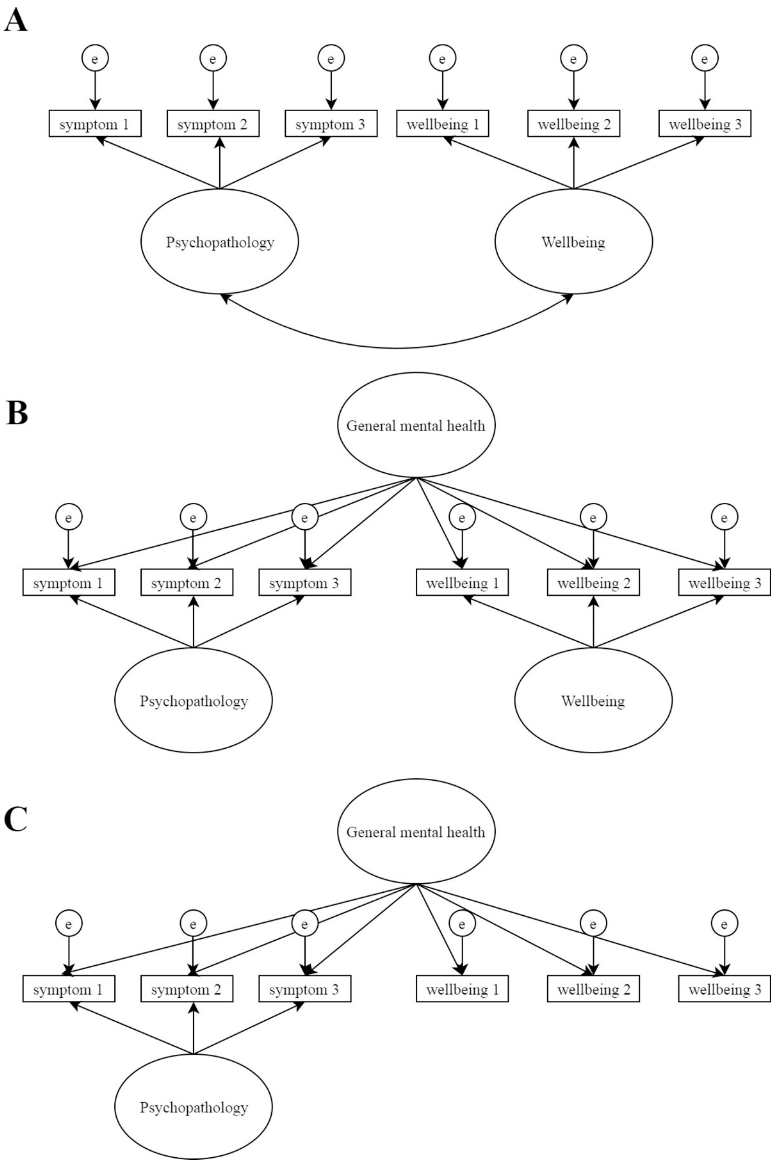


Fig 1. Confirmatory factor analysis model examples. (A) Correlated factors model. (B) Classical bifactor model. (C) *S*-1 model.

https://doi.org/10.1371/journal.pone.0213018.g001

considered, not only because it is statistically more robust than the classical bifactor model, but also because it provides an opportunity to generate an interpretable measurement structure in the presence of general covariance but not essential unidimensionality.

Despite such criticisms, some argue bifactor models can be successfully used when essential unidimensionality is supported, such that the specific factors represent noise (e.g. method factors) [59, 60]. Such a structure was found for mental health difficulties and wellbeing in adults [40], suggesting that this should be tested in adolescence (despite the potential noise introduced by GHQ-12 noted above). Furthermore, bifactor models provide a platform to examine dimensionality via a robust method, the Explained Common Variance (ECV) index [61–63]. Though the question of dimensionality has underpinned much dual-factor research, this has yet to be statistically explored. However, for the reasons described above, and despite common practice [28, 41], we suggest that bifactor structures should not be accepted and interpreted merely based on model fit, especially when unidimensionality is not supported.

It has also been recently pointed out that measurement structures, such as bifactor models, should not be interpreted as evidence of broader construct validity, beyond measures employed [60]. The purpose of this study, however, is to demonstrate an example of models and methods needed, given that mental health difficulties and wellbeing are routinely used together as outcomes in adolescent research [2–5, 13]. We therefore aim to provide evidence of their measurement structure so that bias through failing to account for covariance, can be avoided, rather than to present a definitive structure.

The current study

On the basis of the evidence reviewed above, several predictions were made. Firstly, latent wellbeing would be correlated with latent mental health difficulties factors, particularly internalising, at moderate levels (hypothesis 1). This hypothesis was operationalised in a correlated factors model (see Fig 1A). Secondly, we predicted that a classical bifactor solution (see Fig 1B) would fit the data well, but that this would not be essentially unidimensional as found by Böhnke et al. [40], since we used more clearly dissociated measures, and research with adolescents has also suggested multidimensionality (hypothesis 2) [41]. Thirdly, if hypotheses one and two were supported, we predicted that an $S\text{-}1_{\text{wellbeing}}$ model (see Fig 1C) would provide a useful and robust structure to account for the covariance of mental health difficulties and wellbeing (hypothesis 3). This model would provide an indication of wellbeing corrected for symptoms. Finally, given that group differences have been noted across gender and income for both outcomes, we explored invariance and associations for the strongest model, based on a balance of psychometric rigor, interpretability and fit (hypothesis 4).

Method

We conducted secondary analysis of baseline data from an evaluation of locally developed interventions designed to prevent mental health problems in young people from 12 areas of England (HeadStart) [64]. The University College London Research Ethics Committee granted ethical approval, and parental consent was given for early adolescents to complete the secure online surveys during their usual school day. Teachers read out an information sheet to pupils before these were completed. This emphasised pupils' confidentiality and their right to withdraw.

Participants

A total of 1982 pupils in their final year of primary education (1051 male, 53%) were drawn from 59 schools in England. Pupils' age ranged between 10.75 and 12.25 (M = 11.21, SD = .30). The sample was not drawn to be representative since it reflected the areas participating in the HeadStart programme. As such, statements of special educational needs were below average (1.3% compared to the national average of 2.8%), while those with registered additional needs not meeting the threshold for a statement was above the national average (21.7% compared to 15.4%) [65]. The percentage of participants from white, non-ethnic minority backgrounds was also slightly above the national average for primary schools (74% compared to 70%) [66], while the number of those exposed to a language at home other than English was similar (20% compared to 19%) [66]. In terms of deprivation, 24% of participants were eligible for free school meals (FSM) when data were collected. This is above the national average of 15.6% [66], but typical of U.K. early adolescents' mental research in schools [67].

Measures

Self-report measures (see S1 Appendix) were used since at age 11 these are a valid indication of early adolescents' internal perspectives [68]. Though externalising symptoms can be more accurately reported by a parent or teacher, internalising and wellbeing symptoms are considered to be more reliable from the child's perspective [68]. Given that informant type may have an impact on the modelling structure and therefore act as a confound, the limitation of self-report for externalising was seen to be outweighed by the strength of using a single informant in the specific analysis conducted.

Mental health difficulties. Mental health difficulties was measured through the Me and My School (M&MS; also referred to as Me and My Feelings) questionnaire, which consists of 10 internalising, and six externalising items [69]. This measure was designed to provide a similar screening function to the Strengths and Difficulties Questionnaire [70], but for a younger age range. Participants responded *never*, *sometimes* or *always* (coded one to three) to brief statements (e.g. "I worry a lot"). Possible scores therefore ranged from 10–30 for internalising and 6–18 for externalising, assuming no missing responses. M&MS has been found to be psychometrically robust, with good internal consistency (in 11–12 year-olds, externalising $\alpha = .80$, internalising $\alpha = .77$); concurrent validity, $r = .67$ - .70, for equivalent, and $r = .22$–24 for non-equivalent subscales of the Strengths and Difficulties Questionnaire; and good known-groups validity between clinical and non-clinical populations [71]. M&MS contains one reverse-coded item in the externalising subscale (item 14 "I am calm").

Wellbeing. Wellbeing was measured by the four-item Child Outcome Rating Scale (CORS) [72]. Four aspects (me, school, family and everything) were responded to by clicking on a smooth line between a happy and sad face. For online administration, this line was measured from 0–100, but then divided by 10 for analysis to match the paper version and facilitate model convergence. Possible scores therefore ranged between 0–10 for each item. CORS has been found to be psychometrically robust with good internal consistency ($\alpha = .84$), test-retest reliability ($r = .60$), and concurrent validity (care-taker CORS, $r = .63$, care-taker Youth Outcome Questionnaire, $r = -.43$)[72]. These researchers also found good responsiveness and known-groups validity between clinical and non-clinical samples.

Family income. Pupil FSM eligibility is captured in a number of ways in England [73]. In the current study, data were used on whether pupils had *ever* been eligible for FSM, rather than their *current* status, since transitions in and out of poverty as well as persistent and current poverty, have all been shown to be associated with child and adolescent mental health [50]. Of the sample, 43% ($N = 860$) had ever been eligible for FSM.

Procedure

Survey data were collected in schools in spring 2015 through a secure online portal and subsequently matched to individual socio-demographic characteristics drawn from the National Pupil Database.

Statistical analysis

Confirmatory factor analysis (CFA) was conducted using Weighted Least Squares with Means and Variance adjustment (WLSMV) in Mplus 8.1. One exception to this was the CFA of the CORS instrument, for which robust maximum likelihood was used since all items were continuous. WLSMV was selected to account for the categorical nature of the M&MS measure [74], handle the substantial floor effects associated with screening measures [75], and because this estimator has been shown to produce minimal bias with clustered data [76]. In addition, correlated residuals, which are better handled by WLSMV [77], were of particular interest in the current study given the tendency of the classical bifactor model to absorb unmodeled complexity of this kind [58]. Finally, WLSMV is recommended where there are a large number of variables and factors, and sample size is large [77], as was the case in the current study.

Chi-square statistics are reported but not used to judge fit given their known sensitivity to sample size. The Comparative Fit Index (CFI), Tucker Lewis Index (TLI) and Root Mean Square Error of Approximation (RMSEA), and its 90% confidence interval (CI) are reported to indicate model fit, with values close to .95 for CFI and TLI, and .06 for RMSEA, typically interpreted as good fit [78]. However, given the overfitting problems associated with bifactor solutions, these indices were interpreted alongside the psychometric rigor of each model as well as other indices such as the ECV.

Evaluation of error variances. Given the problems with not modelling correlated systematic error where this is indicated by modification indices and theoretically supported [58, 59], this was investigated in all instruments and solutions before final models were estimated. Individual CFAs of each instrument were therefore conducted in addition to the models shown in Fig 1, so that systematic error could be evaluated here as well. The evaluation of each instrument at this stage also allowed assessment of how well factors were indicated by items, via loadings. In addition to this we calculated Cronbach's α as basic description of subscale reliability to further ensure all items were appropriate for subsequent analysis.

While in a strict sense bifactor modeling assumes zero error covariances, where this error is systematic (e.g. due to similar wording), the question of correlated errors is one that can be tested [79, 80]. Furthermore, while correlating error terms limits the causal power of the latent factor [81], dimensional covariance between measures was of interest in the current study rather than latent disorders. We therefore included correlated error terms in the current analysis, in line with Reise et al. [59].

Evaluation of mental health models. Intra cluster correlations for indicator variables were calculated to assess non-independence due to sampling from schools. Since these were relatively low (.004-.067), clustering was accounted for using the *type = complex* option in Mplus, which adjusts the chi-square statistics and standard errors based on non-independence [82]. After estimating the models described in hypotheses 1–3, these were compared using chi-square difference testing: Each of the correlated factors and *S*-1 models were nested in the bifactor solution following Reise [83].

Explained common variance. ECV represents a ratio of variance explained by the general factor to that explained by the specific factors, while the Percentage of Uncontaminated Correlations (PUC) provides the percentage of correlations that inform on the general factor relative to the specific factors [61]. When PUC is higher (more correlations relate to the general than the specific factors), less bias is introduced by misfitting a unidimensional structure to multidimensional data. High PUC in combination with moderate to high ECV suggests that though a bifactor, multidimensional structure fits well, there is a strong case for modelling the construct as unidimensional. This is because the general factor would account for most of the variance, and factor loadings in a unidimensional model would likely be very similar to those on the general factor [62]. Reise et al. [61] suggest that PUC > .80 and ECV > .60 may be sufficient to consider unidimensionality.

Group differences. Gender and income measurement invariance were tested for the final model through multigroup CFA. To account for the categorical nature of the M&MS items, a three-step procedure was employed: This involved the estimation of baseline models in each subgroup separately; a configural measurement invariance model, where all loading, threshold and intercept parameters were freely estimated in both groups; and a scalar measurement invariance model where loadings and intercepts/thresholds were considered in tandem, and constrained to be equal across groups [84]. Model-based associations between latent mental health factors and gender and income were then explored via individual regression statements, rather than correlations, due to the categorical nature of the exogenous variables income and gender.

Results

Preliminary analysis

Gender was available for every child, ever FSM eligibility was missing for .9% of the sample, while for M&MS and CORS items, missing data ranged from .6–2.6%. Data were assumed to be missing at random, due to absence on the day of data collection, error or omission of individual items, or lack of up-to-date records from the National Pupil Database. The trivial amount of missing data confirmed that results would likely not be negatively affected by using the limited information estimator WLSMV [77].

Descriptive statistics and correlations are presented in Table 1. As expected, observed wellbeing was moderately associated with both observed mental health difficulties domains, though not with gender or family income. Family income was also not significantly associated with internalising. Externalising symptoms were inversely related to being a girl, as expected.

Evaluation of measurement models and correlated error variances

M&MS. Although acceptable internal consistency was found for both M&MS subscales (externalising $\alpha = .776$; internalising $\alpha = .792$), preliminary CFA indicated a poor factor loading for one item ("I am shy", $\lambda = .291$), which was consistent with other analyses [28, 69]. This item also had a low item total correlation ($r = .257$), and its removal improved internal consistency ($\alpha = .799$). Furthermore, we felt this item could be interpreted as conceptually different from the others (see S1 Appendix), as it is the only one clearly linked to social functioning. The fit of the initial two-factor M&MS scale, χ^2 (103) = 549.444, $p < .001$, RMSEA = .047 (90% CI = .043-.051), CFI = .955, TLI = .947, remained good following this item's removal, χ^2 (89) = 511.309, $p < .001$, RMSEA = .049 (90% CI = .045-.053), CFI = .958, TLI = .951.

Modification indices supported three pairs of correlated residuals between items with similar conceptual content and or wording. These were M&MS items 1 and 3: "I feel lonely" with "Nobody likes me"; M&MS items 5 and 6: "I worry when I am at school" with "I worry a lot"; and M&MS items 7 and 8 "I have problems sleeping" with "I wake up in the night". The inclusion of these correlated error terms resulted in good model fit, χ^2 (86) = 262.342, $p < .001$, RMSEA = .032 (90% CI = .028-.037), CFI = .983, TLI = .979, so this modified structure was taken forward.

Table 1. Descriptive statistics and bivariate correlations.

Variable	1.	2.	3.	4.	5.	*M*	*SD*	Min-Max
1. Internalising	–					13.87	3.36	2–27
2. Externalising	.441*	–				8.99	2.46	1–18
3. Wellbeing	-.439*	-.329*	–			32.40	7.31	0–40
4. Gender[a]	.087*	-.149*	.026	–				
5. Income[b]	.034	.150*	-.036	.033	–			

[a] 0 = boys, 1 = girls

[b] 0 = never eligible for free school meals, 1 = ever eligible for free school meals.

* $p < .01$.

https://doi.org/10.1371/journal.pone.0213018.t001

Table 2. Fit of confirmatory factor analysis models.

Model	χ^2 (df)	RMSEA(90% confidence interval)	CFI	TLI	χ^2difference (df)
1. Correlated Factors	410.931** (145)	.030 (.027, .034)	.972	.967	-
2. Bifactor	321.561**(129)	.027 (.024, .031)	.980	.973	1. vs. 2. 110.742**(16)
3. S-1$_{\text{wellbeing}}$	535.155**(133)	.039 (.036, .043)	.958	.946	2. vs. 3. 187.072**(4)
4. S-1$_{\text{internalising}}$	407.180**(138)	.031 (.028, .035)	.972	.965	2. vs. 4. 87.311**(9)

RMSEA, Root Mean Square Error of Approximation; CFI, Comparative Fit Index; TLI, Tucker Lewis Index.

** $p < .001$.

https://doi.org/10.1371/journal.pone.0213018.t002

CORS. While internal consistency for CORS was acceptable (α = .745), the model fit of a unidimensional structure was poor, χ^2 (2) = 24.831, $p < .001$, RMSEA = .076 (90% CI = .51-.104), CFI = .976, TLI = .928. Modification indices supported the inclusion of one pair of correlated errors due to conceptual and wording overlap: CORS items 1 and 3 "how am I doing" with "how am I doing at school". The inclusion of this error correlation substantially improved fit, χ^2 (1) = 1.281, p = .258, RMSEA = .012 (90% CI = .000- .062), CFI = 1, TLI = .998, and was therefore taken forward.

Dual-factor mental health models

Hypothesis 1 was supported since the correlated factors model had excellent fit to the data (See Table 2), and significant loadings for all items (λ ≥ .43, see Fig 2). Furthermore, the estimated correlation between latent internalising and wellbeing was found to equal that between the two latent mental health difficulties dimensions (r = -.58). Latent externalising was also found to be substantially related to latent wellbeing, though to a lesser degree than was internalising (r = -.42).

Although these clear relationships were found between constructs, a unidimensional structure was not supported, as predicted in hypothesis two (PUC = .67, ECV = .55). The classical bifactor model did, however, show excellent fit to the data (see Table 2), and each item had at least one salient loading on the general or specific factor (see Fig 3). In addition to the lack of unidimensionality, inspection of the parameter estimates revealed further problems. Four internalising items had very low loadings on the specific factor (unhappy λ = .28; unliked λ = .15; sleep problems λ = .18; wakeup λ = .08), and the factor variance for internalising was also low compared to the externalising factor, which was on the same response scale (ξ = .13 versus ξ = .36). While it could be argued that internalising acted as a particularly good indicator of the general factor, we interpret this result in line with Eid et al. [56], and suggest that this is

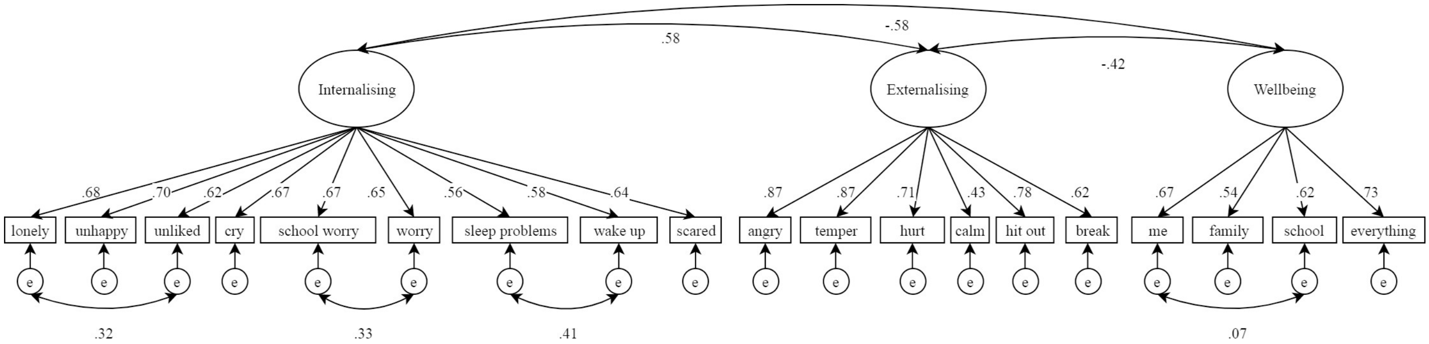


Fig 2. Correlated factors model results.

https://doi.org/10.1371/journal.pone.0213018.g002

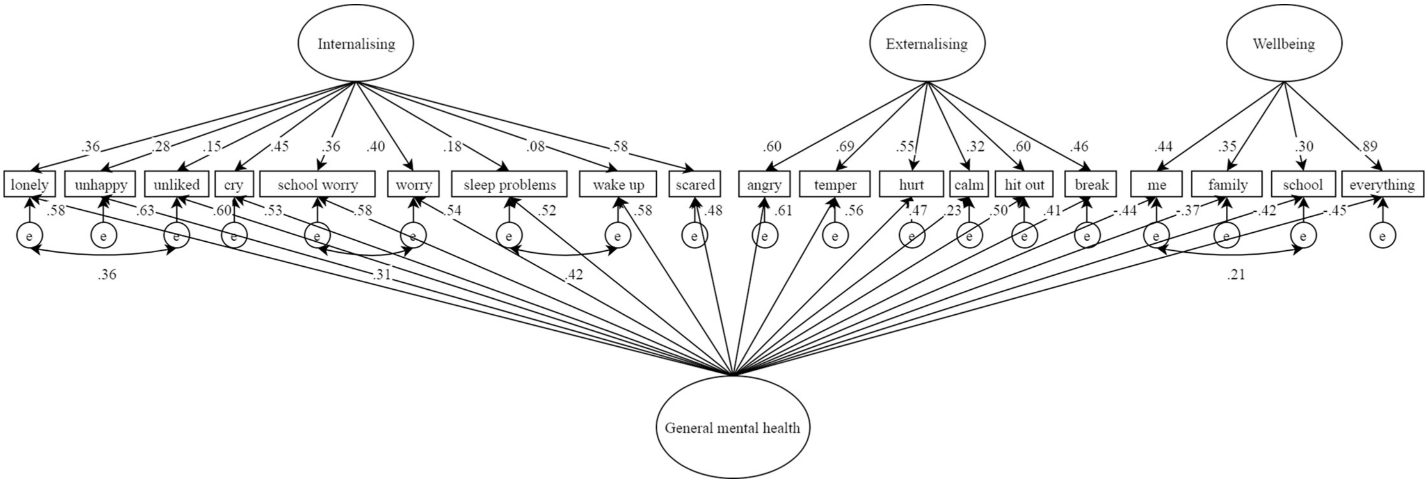


Fig 3. Classical bifactor model results.

https://doi.org/10.1371/journal.pone.0213018.g003

evidence of a *vanishing* factor, a result identified as consistent with the psychometric misspecification of classical bifactor solutions. Though the classical bifactor model therefore showed superior fit to other models estimated, it was rejected based on the ECV and disappearing internalising factor.

Contrary to hypothesis 3, the S-$1_{\text{wellbeing}}$ model was also rejected for a number of reasons. It showed inferior fit compared to the correlated model (which was less likely to overfit), the internalising factor remained relatively weak, consistent with the classical bifactor model, and the general wellbeing factor was more strongly defined by internalising than wellbeing items (see Fig 4). This suggested that general wellbeing covariance in mental health difficulties items was not a good representation of the data. In light of this, and the vanishing internalising factor found in the classical bifactor solution, post-hoc analysis of an S-$1_{\text{internalising}}$ model was conducted (see Fig 5). This model showed almost identical fit to the correlated factors model (see Table 2) and unlike the S-$1_{\text{wellbeing}}$ model, the general factor was this time most strongly defined by its unique items. The general factor in S-$1_{\text{internalising}}$ can therefore be interpreted as

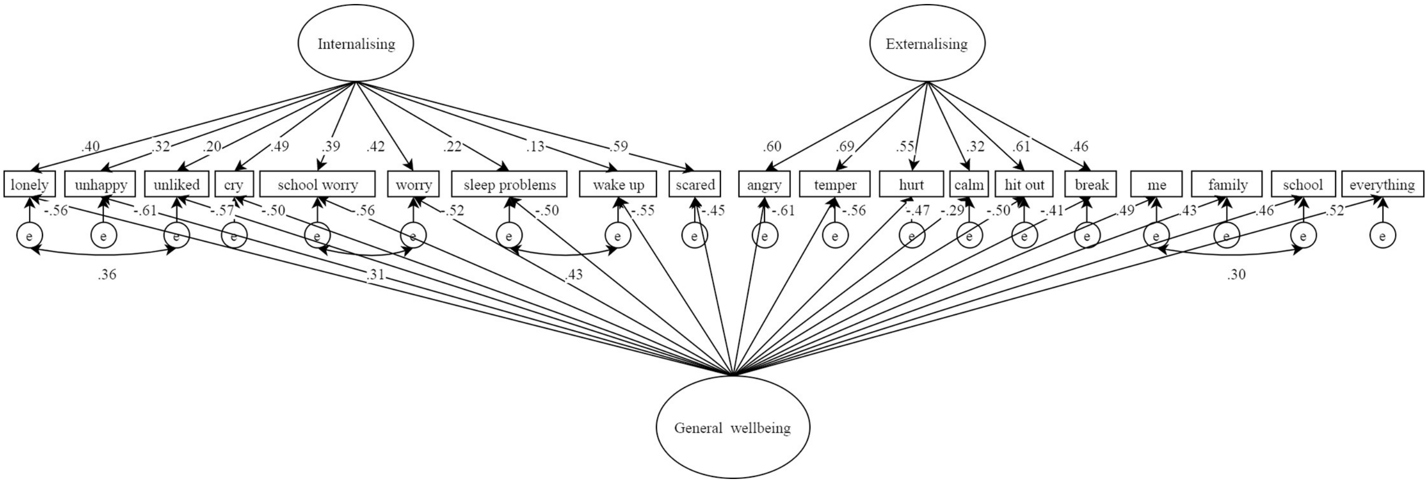


Fig 4. S-$1_{\text{wellbeing}}$ model results.

https://doi.org/10.1371/journal.pone.0213018.g004

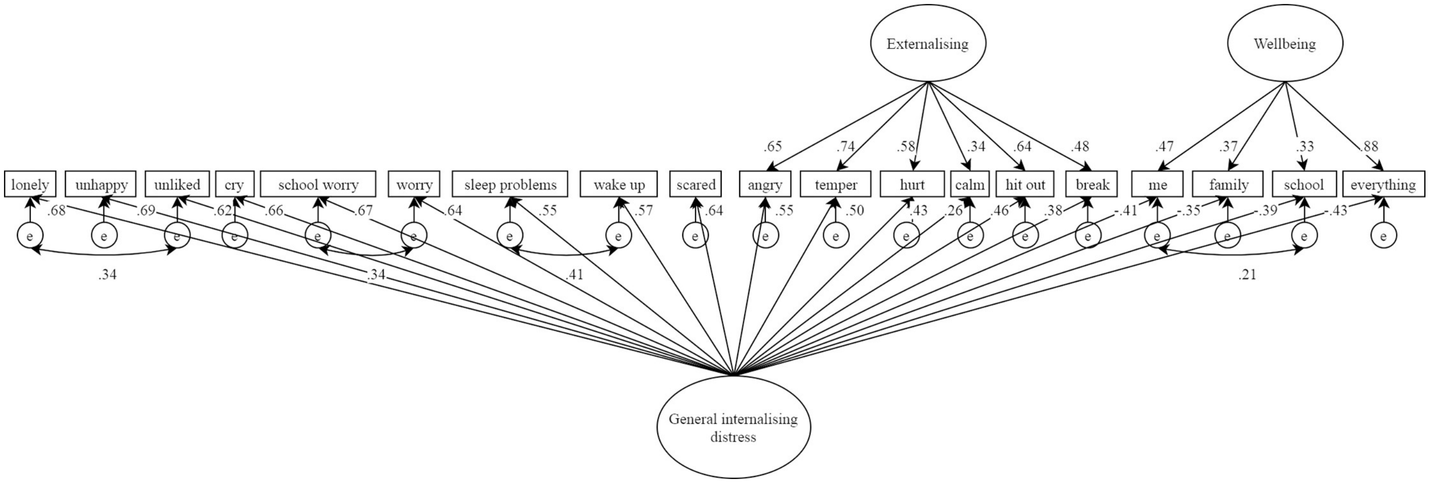


Fig 5. S-1$_{\text{internalising}}$ model results.

https://doi.org/10.1371/journal.pone.0213018.g005

modelling general internalising distress (GID) that is tapped not only by items designed to do so, but also variance of this construct captured by externalising and wellbeing items.

Difference testing was conducted between models where possible (based on number of parameters and the Nesting and Equivalence Test, NET) [85]. Of the possible comparisons, the classical bifactor model was the best as expected. It has been suggested that comparisons between models of the types we explored here should be interpreted with caution due to mathematical closeness [57]. Indeed, fit statistics revealed the correlated factors and S-1$_{\text{internalising}}$ models to be extremely similar, though the latter appeared to be slightly worse based on qualitative inspection of fit statistics (this was necessary since the NET procedure revealed these models were not nested). Though the correlated factors model was therefore likely the best given its relative parsimony [74], and we recommend it be retained where possible in similar analysis, hypothesis 4 was considered in both correlated factors and S-1$_{\text{internalising}}$ models since each are useful for different scenarios (see discussion below).

Measurement invariance testing. Invariance testing was therefore conducted on both of these models and results can be seen in Table 3. Partial measurement invariance was supported for gender in both models, with the items "I cry a lot" showing non-invariance in both, and the item "How am I doing at school" showing non-invariance in the correlated factors model. Full measurement invariance was supported for income in both models, though a small negative residual variance (-.14) was found for CORS4 ("How is everything going?") in the ever FSM group for the S-1$_{\text{internalising}}$ model. This impossible result appeared to arise from the correlated error term between the CORS items "How am I doing?" and "How am I doing at school?", which was retained in the model since it was significant and meaningful, $r = .26$. In line with Muthén [86], the residual variance of CORS4 was fixed to zero since this parameter was non-significant ($p = .84$), and fixing this to zero did not substantially change the model fit. Since full measurement invariance is frequently seen to be untenable [87], we interpreted these results as indicating that models functioned reasonably well across the groups studied.

In order to estimate the association of latent mental health factors with gender and income, non-invariant items were removed from both correlated factors and S-1$_{\text{internalising}}$ models [88–90] . Their removal resulted in slightly better fitting models (correlated factors without non-invariant items, $\chi^2 = 311.847^*(113)$, RMSEA = .030, (90% CI = .026-.034) CFI = .978, TLI = .967; S-1$_{\text{internalising}}$ without non-invariant item, $\chi^2 = 364.857^*(121)$; RMSEA = .032 (90% CI = .028-.036); CFI = .973; TLI = .966) possibly due to removal of noise, and or the fact that CFI is

Table 3. Results of multigroup invariance testing.

Model	χ^2 (df)	RMSEA (90% confidence interval)	CFI	TLI	χ^2difference (df)
Correlated Factors gender invariance					
Boys baseline	295.292**(145)	.031 (.026, .037)	.970	.965	-
Girls baseline	258.267**(145)	.029 (.023, .035)	.980	.976	-
Configural	740.155**(298)	.039 (.035, .042)	.958	.952	-
Scalar	736.419**(345)	.034 (.030, .037)	.963	.964	82.734**(47)
Scalar M&MS4/CORS3 free	714.075**(340)	.033 (.030, .037)	.965	.965	56.103 (42), p = .07
***S*-1 gender invariance**					
Boys baseline	294.634**(138)	.033 (.028, .038)	.969	.962	-
Girls baseline	242.902**(138)	.029 (.023, .034)	.981	.977	-
Configural	746.480**(284)	.041 (.037, .044)	.957	.948	
Scalar	712.306**(343)	.033 (.030, .036)	.965	.965	94.405** (59)
Scalar M&MS4 free	695.876**(340)	.032 (.029, .036)	.967	.966	67.778 (54), p = .10
Correlated factors income invariance					
everfsm baseline	274.547**(145)	.032 (.026, .038)	.975	.971	-
neverfsm baseline	281.287**(145)	.029 (.024, .034)	.970	.964	-
Configural	749.155**(298)	.039 (.036, .043)	.953	.946	-
Scalar	698.214**(345)	.032 (.029, .036)	.963	.964	44.060 (47), p = .60
***S*-1 income invariance**					
everfsm baseline	268.413**(139)	.033 (.027, .039)	.975	.969	-
neverfsm baseline	274.571**(138)	.030 (.025, .035)	.970	.962	-
Configural	769.994**(284)	.042 (.038, .045)	.950	.939	-
Scalar	672.437**(341)	.031 (.028, .035)	.966	.966	50.095(57), p = .73

RMSEA, Root Mean Square Error of Approximation; CFI, Comparative Fit Index; TLI, Tucker Lewis Index; M&MS4, "I cry a lot"; CORS3, "How am I doing at school".

** $p < .001$.

https://doi.org/10.1371/journal.pone.0213018.t003

known to be sensitive to the number of items [91]. For both models, wellbeing was not significantly associated with gender, internalising was modestly associated with being a girl, and externalising was substantially associated with being a boy (see Table 4). In line with the observed score correlations in Table 1, only externalising was significantly associated with low family income in either the correlated factors or S-$1_{\text{internalising}}$ models.

Discussion

The aim of the current study was to further our understanding of the structure of mental health difficulties and wellbeing in early adolescence, using secondary data from a large U.K. sample (N = 1982). Despite existing theoretical frameworks (e.g., two-continua approach) [39], the robust analysis of the measurement structure of mental health difficulties and

Table 4. Gender and income associations with mental health factors.

Correlate	Internalising		Externalising		Wellbeing	
	M1	M4 (GID)	M1	M4	M1	M4
Gender	.192*	.173*	-.375*	-.612*	-.041	.116*
Income	.080	.082	.363*	.393*	-.077	-.030

M1, correlated factors model; M4, S-$1_{\text{internalising}}$; GID, general internalising distress.

* $p < .01$.

https://doi.org/10.1371/journal.pone.0213018.t004

wellbeing, and especially in younger populations, has been lacking from the extant literature. Given recent limitations pertaining to common methodological approaches, such as bifactor modeling [56–59], alternative methodologies were considered (ECV, S-1), and competing CFA models were estimated, which allowed for a more robust representation of the comprehensive mental health model.

Overall, unidimensionality was not supported in the current study. Instead, our results demonstrate that mental health difficulties and wellbeing are distinct but related constructs and should therefore be considered alongside each other within late childhood-early adolescent research. The simple correlated factors structure fitted the data well and revealed that wellbeing was just as related to internalising difficulties as this was to externalising symptoms. Despite the superior fit of the bifactor model, this was rejected in the current study, as the general factor explained only 55% of the total common variance. Results from the S-1 models further revealed that a general internalising distress factor could play an important role in all item responses. Partial gender and full income measurement invariance were established for the correlated and $S\text{-}1_{\text{internalizing}}$ models. However, given that the correlated model was the most parsimonious, with a slightly better fit than that of $S\text{-}1_{\text{internalizing}}$, we considered that to be the most theoretically and statistically plausible model of comprehensive mental health.

In line with previous findings [38], medium to large latent correlations were observed between wellbeing and mental health difficulties domains. The present study, however, accounted for the known distinction between childhood internalising and externalising symptoms [47], rather than conflating these as has sometimes been the case [7]. This also enabled comparison of effect sizes for estimated correlations between all latent constructs in the correlated factors model and demonstrated that wellbeing was no more dissociated from mental health difficulties constructs than these were from one another. This strengthens the idea that wellbeing may be used to *calibrate* psychopathology scores [40], and provides clear justification for the inclusion of wellbeing in mental health models.

In contrast to previous research [28, 41], we did not accept the classical bifactor solution as the final model, despite its superior fit. Since the general factor explained only 55% of the total common variance, the classical bifactor model was substantively uninterpretable, and was therefore rejected. In other words, while some previous research has suggested symptoms of mental health difficulties and wellbeing could be considered a single continuum [40], in line with hypothesis 2 our findings did not support this. We found that when internalising, externalising and wellbeing were modelled together in a large sample of early adolescents, these constructs should be treated as distinct but related factors. As suggested earlier, our choice of M&MS as a mental health difficulties measure capturing more than just negative affect, and the age of our sample, are likely to have contributed to our contrasting results. It should also be noted that this lack of support for unidimensionality is somewhat consistent with research with older adolescents [41], though in contrast to this work, we followed recent criticisms and rejected the multidimensional bifactor solution [56, 60]. This was in part facilitated by our inclusion of the ECV, which had not been considered in mental health difficulties and wellbeing bifactor models previously, and reinforces the importance of not solely relying on model fit.

Insights from stochastic measurement theory also allowed models with better defined factors to be estimated [56]. Though our hypothesised $S\text{-}1_{\text{wellbeing}}$ model presented a poor fit, parameter estimates in the classical bifactor solution led to post-hoc analysis of an $S\text{-}1_{\text{internalising}}$ model which explained the data well. This post-hoc analysis was conducted since internalising appeared to be weakened as a specific factor in the classical bifactor and $S\text{-}1_{\text{wellbeing}}$ solutions, but showed strong loadings on the general factors in both models. In line with Eid et al. [56], we therefore considered a model in which specific internalising was removed, allowing internalising items to define the general factor. Since relatively stable general loadings were also observed across the classical bifactor and both *S*-1 models, GID covariance may have been responsible for each of these models' general factors. Moreover, in the $S\text{-}1_{\text{wellbeing}}$ model the strongest loadings on the general factor were seen for internalising, rather than wellbeing items as would be expected. Statistical comparison was not possible between the correlated factors and $S\text{-}1_{\text{internalising}}$ models, and in fact it has been suggested anyway that comparison of such models is problematic, due to their mathematical closeness [57]. Nevertheless, the correlated model appeared to have slightly better fit than the $S\text{-}1_{\text{internalising}}$ model, and since this was the simpler solution, we suggest that this should be preferred where possible.

This is not say, however, that the $S\text{-}1_{\text{internalising}}$ model is inadmissible, as such a model would be able to address certain research questions unanswerable by the correlated factors solution. For instance, where the specific role of external correlates is of interest for particular mental health domains, as explored by Patalay et al. [7], $S\text{-}1_{\text{internalising}}$ would allow researchers to estimate the effects of these on GID, externalising behaviour and wellbeing separately, while controlling for each of the other outcomes. While $S\text{-}1_{\text{internalising}}$ was considered less optimal, particularly since it had more parameters, in combination with the other models and ECV results, it provides further insight into previous research. For this reason, our discussion focuses on the interpretation of both the correlated and $\text{S-}1_{\text{internalising}}$ models.

For instance, together, our models shed light on previous findings relating to internalising. Specifically, externalising and wellbeing group factors have tended to show substantial loadings after accounting for a general factor, whereas internalising loadings have behaved differently, becoming small, sometimes insignificant, and even negative on occasion [28, 40, 41]. The $S\text{-}1_{\text{internalising}}$ model could clarify this since it represents the influence of a latent internalising trait on responses to all mental health difficulties and wellbeing items. Such a structure could therefore underlie other bifactor solutions, since the consistent presence of relatively weak specific internalising suggests that this could be defining other general factors found [28, 40, 41, 56].

Theoretically GID is also consistent with the wider literature, since some of the covariance with wellbeing could be explained by the conceptual overlap (e.g. happiness and unhappiness). Covariance with externalising, on the other hand could reflect known comorbidity, which is thought to arise for a number of complex reasons, including method factors as well as cascading or predisposing effects [20, 92, 93]. Previous research has often combined internalising and externalising symptoms when considering the relationship of mental health difficulties to wellbeing [1, 3, 13]. However, our study suggests this may be problematic since both overlap and dissociation between constructs was found. It is possible that overlap at the latent level explains response patterns, and that dimensions such as those we propose should be considered rather than summed scores. While some research has categorised young people according to flourishing, languishing, etc., latent dimensional approaches could yield different results. For instance, in the $S\text{-}1_{\text{internalising}}$ model it is possible that those with considerable GID show tendencies towards languishing, while those with behavioural externalising symptoms, separate from distress, could show higher wellbeing. A symptomatic but content group could therefore arise under circumstances in which the behavioural aspect of externalising is tapped as psychopathology in early adolescents who are not distressed, and therefore in turn report high wellbeing.

The estimation of both *S*-1 models in the current study, in combination with the calculation of ECV in the bifactor model, clarified the covariance structure of the items. This is namely that just over half of all common variance could be explained by a classical general factor, but that this is likely due to shared internalising variance across all items. While the current study draws on a relatively new area of work [56], current findings support the wider utility of *S*-1 models. These have not only addressed some of the concerns raised around bifactor modeling [56, 60], but also added substantive theoretical insight.

Having explored the covariance structure of mental health domains, our final aim was to shed light on their complex relationships with gender and family income. Externalising symptoms are often associated with boys, and emphasis tends to be on girls reporting higher internalising symptoms because of elevated rates in later adolescence [6, 49]. However, there is evidence that internalising symptoms also play an important role in boys' psychopathology and externalising symptoms [67, 93]. For instance, initial lower levels of internalising were shown to predict lower levels of externalising at a later time point in both boys and girls [67].

Consistent with these studies, our results suggest only a weak association of internalising distress with gender in early adolescence. For both the correlated and S-$1_{\text{internalising}}$ models internalising (at the specific level for the former, and global GID level for the latter) showed a small association with being a girl. Therefore, when specific externalising behaviour (not associated with GID) was accounted for in the S-$1_{\text{internalising}}$ model, girls still showed only slightly higher levels of GID than boys. Similarly when the effect of latent internalising on externalising item responses was accounted for, the association of being a boy with externalising behaviour was notably much larger. This therefore suggests that while behavioural problems were associated with being male, this was particularly the case after controlling for GID. Furthermore, when poor behaviour (not associated with distress) was accounted for, girls still showed only slightly higher levels of internalising distress than boys. An alternative explanation for this finding could be that externalising psychopathology is entirely distinct from internalising, and remained associated with being a boy for this reason. However, five of the six externalising items had salient loadings on the GID factor (λ = .38-.64), suggesting that these items were well defined by GID, and these constructs were therefore not entirely separate.

As with gender, the associations found in the current study between mental health factors and income advance previous work which treated these factors as a single variable [7]. It was unsurprising that wellbeing did not show significant associations with low income [10]. However, it was more unexpected that only externalising was significantly and substantially related to this outcome [7], though similar conduct and emotional domains have shown stronger associations to income for the former than the latter [50]. The discrepancy in significance may therefore be due to the use of a larger sample by Fitzsimons et al. [50].

Beyond the benefits of adding S-1 models to understand covariance and relationships to key outcomes, the modeling approach was also strengthened by the inclusion of correlated errors. These were included to avoid overfitting in an entirely locally independent bifactor model, such that covariance beyond specific latent constructs would be absorbed by the general factor [58, 59]. These were carefully evaluated according to item content, wording and modification indices. Though inclusion of such parameters weakens the causal power of the latent trait, it is untenable to assume no relationship between conceptually similar items such as "I have problems sleeping" and "I wake up in the night" [81]. While CFA was used, the current study was somewhat exploratory, investigating the dimensionality of mental health difficulties and wellbeing, therefore allowing for relationships beyond hypothesised factors. In addition, consistent with recent calls [34], our analysis was focused at a symptom level. It therefore did not assume causal disorders, but rather considered the covariance structure of items. Nevertheless, it remains important to understand that there are associations between items beyond the latent traits modelled. As stated previously, the analysis of comprehensive mental health put forward here is not an attempt to conceptualise a definitive structure of "positive" and "ill" mental health. If such an approach were adopted, the violation of local independence would be potentially more serious in our view. Rather, our hypotheses, findings and discussion were designed to interrogate measurement assumptions routinely made for these outcomes in research with young people.

It is clear that epidemiological measures, such as those used here, can be problematic in terms of item content for local independence assumptions. While some would argue that alternative approaches to latent trait models should therefore be adopted, we feel that the robust analysis of dimensionality and covariance provided here was a key first step, before further exploration or alternative approaches considering mental health difficulties and wellbeing items together could be employed. If strong relationships between constructs had not been found in the present analysis, there would be little value in further study. It could be argued that analysis of the kind we have presented should have been employed even sooner, before analysis of correlates was considered. Our critical review of the literature and findings also suggest that categorical treatment of these outcomes can be problematic, and does not appear to be a good representation of the data. This reinforces that previous treatment of the outcomes as such [1–5, 13, 14] may lead to false conclusions.

However, it should be noted that the latent trait account we have offered may not be the only reason items covaried as they did, and that other approaches such as network analysis should be considered in future [94]. It has also been demonstrated that complex bifactor solutions can overfit data when these account for unusual response patterns [59]. Estimating the percentage of respondents who fit the model to ascertain whether complex solutions account for a minority implausible response patterns as Reise et al. [59] did, would also be pertinent to dual-factor research, given the consistent finding that a minority are neither flouring nor languishing [1–5, 13, 14].

This was the first study to our knowledge to empirically explore the structure of latent mental health difficulties and wellbeing in early adolescence. Furthermore, we employed more appropriate measures and robust approaches to bifactor modelling than those commonly used [40, 41]. Unidimensionality was not supported, but clear justification was found for the inclusion of wellbeing in mental health models, and GID was found to explain responses to all items at a salient level. This study therefore draws together and improves on school psychology dual-factor [1–5, 13, 14], and mental health bifactor research [27, 28, 30]. While the former has tended to categorically dichotomise mental health difficulties and wellbeing, and therefore lose important information [34], the latter has generally failed to account for the statistical properties of bifactor models, leading to potentially misleading conclusions [56].

Despite the use of rigorous methodology, several limitations should also be acknowledged. Firstly, the exploration of any construct is tied to the measures used, and results will inevitably vary by instrument, as already seen in the contrast between the present study and that by Böhnke et al. [40]. Though well-validated instruments were selected, replication studies should consider employing alternative measures. Similarly, constructs were assessed via self-report measures for feasibility and design reasons and as already noted, externalizing symptoms may be more accurate when reported by an adult. However, wellbeing and internalizing symptoms are likely more valid from the young person's perspective [68]. Informant reports are also limited in that the informant (e.g. parent, teacher) typically only observes the adolescent in a single context [95]. Use of mixed informants would also likely have acted as a confound since self and informant ratings are often only weakly or moderately correlated, particularly for children and adolescents [96–98]. Though the sample size was substantial and met the recommended minimum *N*:*q* ratio (at 25.7:1), future research, particularly if more complex structural predictive components are added, should consider Monte Carlo simulations for decisions on sample size [99]. The representativeness of the sample may also be considered a limitation since poorer adolescents were overrepresented, though as stated previously, rates here were comparable to other U.K. school-based mental health research. FSM eligibility has also been criticised as a measure of socioeconomic status and proxy for family income [100], and though efforts were made to mitigate this through the use of everFSM, future studies should consider including more accurate and comprehensive measures of family income. Finally, this study used the relatively new ECV and PUC indices. While some thresholds have been recommended for these [61], further research is needed to confirm their accuracy.

Conclusion

In the first study of its kind, early adolescents' comprehensive mental health was explored using a large sample and robust analytical strategy. Previous research in mental health and school psychology has been extended, with our results clarifying how general factors may arise, through thorough investigation via the ECV and *S*-1 models. Clear correspondence was found between internalising and externalising symptoms, and wellbeing, and evidence suggested common GID variance was meaningfully predictive of responses to all items. This research therefore offers insight into comorbidity and dual-factor response patterns, since it suggests that common internalising may contribute across mental health domains. Given the problems with bifactor modelling in previous research, and categorical approaches often taken, our analysis provides the first robust platform from which relationships between wellbeing and mental health difficulties domains can be explored further.

Supporting information

S1 Appendix. Items of Me and My School and Child Outcome Rating Scale questionnaires. (DOCX)

Acknowledgments

The data used in this study were collected as part of the HeadStart learning programme. The authors are therefore grateful for the work of the wider research teams at the Anna Freud Centre and the University of Manchester for their role in coordinating the evaluation, as well as collecting and managing the data. The authors also acknowledge the National Pupil Database from which demographic data were obtained. Finally, we are extremely grateful to all students who took part in this study, as well as the local authorities and schools for their help in recruiting them.

Author Contributions

Conceptualization: Louise Black, Margarita Panayiotou.

Formal analysis: Louise Black.

Funding acquisition: Neil Humphrey.

Investigation: Neil Humphrey.

Methodology: Louise Black, Margarita Panayiotou.

Project administration: Neil Humphrey.

Resources: Neil Humphrey.

Supervision: Margarita Panayiotou, Neil Humphrey.

Visualization: Louise Black.

Writing – original draft: Louise Black.

Writing – review & editing: Margarita Panayiotou, Neil Humphrey.

References

1. Antaramian SP, Scott Huebner E, Hills KJ, Valois RF. A Dual-Factor Model of Mental Health: Toward a More Comprehensive Understanding of Youth Functioning. American Journal of Orthopsychiatry. 2010; 80(4):462–72. https://doi.org/10.1111/j.1939-0025.2010.01049.x PMID: 20950287
2. Greenspoon PJ, Saklofske DH. Toward an Integration of Subjective Well-Being and Psychopathology. Social Indicators Research. 2001; 54(1):81–108.
3. Lyons MD, Huebner ES, Hills KJ. The Dual-Factor Model of Mental Health: A Short-Term Longitudinal Study of School-Related Outcomes. Social Indicators Research. 2013; 114(2):549–65.
4. Suldo S, Thalji A, Ferron J. Longitudinal academic outcomes predicted by early adolescents' subjective well-being, psychopathology, and mental health status yielded from a dual factor model. The Journal of Positive Psychology. 2011; 6(1):17–30.
5. Suldo S, Thalji-Raitano A, Kiefer SM, Ferron JM. Conceptualizing High School Students' Mental Health Through a Dual-Factor Model. School Psychology Review. 2016; 45(4):434–57.
6. Patalay P, Fitzsimons E. Mental ill-health among children of the new century: trends across childhood with a focus on age 14. September 2017. London: Centre for Longitudinal Studies; 2017.
7. Patalay P, Fitzsimons E. Correlates of Mental Illness and Wellbeing in Children: Are They the Same? Results From the UK Millennium Cohort Study. J Am Acad Child Adolesc Psychiatry. 2016; 55 (9):771–83. https://doi.org/10.1016/j.jaac.2016.05.019 PMID: 27566118
8. Kovess-Masfety V, Husky MM, Keyes K, Hamilton A, Pez O, Bitfoi A, et al. Comparing the prevalence of mental health problems in children 6–11 across Europe. Social Psychiatry and Psychiatric Epidemiology. 2016; 51(8):1093–103. https://doi.org/10.1007/s00127-016-1253-0 PMID: 27314494
9. Polanczyk GV, Salum GA, Sugaya LS, Caye A, Rohde LA. Annual Research Review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. Journal of Child Psychology and Psychiatry. 2015; 56(3):345–65. https://doi.org/10.1111/jcpp.12381 PMID: 25649325
10. Bradshaw J, Rees G. Exploring national variations in child subjective well-being. Children and Youth Services Review. 2017; 80:3–14.
11. Dinisman T, Ben-Arieh A. The Characteristics of Children's Subjective Well-Being. Social Indicators Research. 2016; 126(2):555–69.
12. Pople L, Society TCs, Rees G. <the-good-childhood-report-2017_full-report_0.pdf>. 2017. p. 1–64.
13. Lyons MD, Huebner ES, Hills KJ, Shinkareva SV. The Dual-Factor Model of Mental Health:Further Study of the Determinants of Group Differences. Canadian Journal of School Psychology. 2012; 27 (2):183–96.
14. Suldo S, Shaffer EJ. Looking Beyond Psychopathology: The Dual-Factor Model of Mental Health in Youth. School Psychology Review. 2008; 37(1):52–68.
15. Jones PB. Adult mental health disorders and their age at onset. British Journal of Psychiatry. 2013; 202(s54):s5–s10.
16. Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. The Lancet Child & Adolescent Health. 2018; 2(3):223–8.
17. Patton GC, Sawyer SM, Santelli JS, Ross DA, Afifi R, Allen NB, et al. Our future: a Lancet commission on adolescent health and wellbeing. The Lancet. 2016; 387(10036):2423–78.
18. Dahl RE, Allen NB, Wilbrecht L, Suleiman AB. Importance of investing in adolescence from a developmental science perspective. Nature. 2018; 554:441. https://doi.org/10.1038/nature25770 PMID: 29469094
19. Kinderman P, Sellwood W, Tai S. Policy implications of a psychological model of mental disorder. Journal of Mental Health. 2009; 17(1):93–103.
20. Krueger RF, Markon KE. Reinterpreting Comorbidity: A Model-Based Approach to Understanding and Classifying Psychopathology. Annual Review of Clinical Psychology. 2006; 2(1):111–33.
21. Carragher N, Krueger RF, Eaton NR, Slade T. Disorders without borders: current and future directions in the meta-structure of mental disorders. Social Psychiatry and Psychiatric Epidemiology. 2015; 50 (3):339–50. https://doi.org/10.1007/s00127-014-1004-z PMID: 25557024
22. Moncrieff J, Timimi S. The social and cultural construction of psychiatric knowledge: an analysis of NICE guidelines on depression and ADHD. Anthropology & Medicine. 2013; 20(1):59–71.
23. Sondeijker FEPL Ferdinand RF, Oldehinkel AJ, Veenstra R, De Winter AF, Ormel J, et al. Classes of adolescents with disruptive behaviors in a general population sample. Social Psychiatry and Psychiatric Epidemiology. 2005; 40(11):931–8. https://doi.org/10.1007/s00127-005-0970-6 PMID: 16222441

24. van Lier PAC, Verhulst FC, van der Ende J, Crijnen AAM. Classes of disruptive behaviour in a sample of young elementary school children. Journal of Child Psychology and Psychiatry. 2003; 44(3):377–87. PMID: 12635967

25. de Nijs PFA, van Lier PAC, Verhulst FC, Ferdinand RF. Classes of Disruptive Behavior Problems in Referred Adolescents. Psychopathology. 2007; 40(6):440–5. https://doi.org/10.1159/000107428 PMID: 17709974

26. Forbes MK, Tackett JL, Markon KE, Krueger RF. Beyond comorbidity: Toward a dimensional and hierarchical approach to understanding psychopathology across the life span. Development and Psychopathology. 2016; 28(4pt1):971–86. https://doi.org/10.1017/S0954579416000651 PMID: 27739384

27. Caspi A, Houts RM, Belsky DW, Goldman-Mellor SJ, Harrington H, Israel S, et al. The p Factor:One General Psychopathology Factor in the Structure of Psychiatric Disorders? Clinical Psychological Science. 2014; 2(2):119–37. https://doi.org/10.1177/2167702613497473 PMID: 25360393

28. Patalay P, Fonagy P, Deighton J, Belsky J, Vostanis P, Wolpert M. A general psychopathology factor in early adolescence. Br J Psychiatry. 2015; 207(1):15–22. https://doi.org/10.1192/bjp.bp.114.149591 PMID: 25906794

29. Carragher N, Teesson M, Sunderland M, Newton NC, Krueger RF, Conrod PJ, et al. The structure of adolescent psychopathology: a symptom-level analysis. Psychological Medicine. 2015; 46(5):981–94. https://doi.org/10.1017/S0033291715002470 PMID: 26620582

30. Castellanos-Ryan N, Brière FN, O'Leary-Barrett M, Banaschewski T, Bokde A, Bromberg U, et al. The structure of psychopathology in adolescence and its common personality and cognitive correlates. Journal of Abnormal Psychology. 2016; 125(8):1039–52. https://doi.org/10.1037/abn0000193 PMID: 27819466

31. Tackett JL, Lahey BB, van Hulle C, Waldman I, Krueger RF, Rathouz PJ. Common genetic influences on negative emotionality and a general psychopathology factor in childhood and adolescence. Journal of Abnormal Psychology. 2013; 122(4):1142–53. https://doi.org/10.1037/a0034151 PMID: 24364617

32. Waldman ID, Poore HE, van Hulle C, Rathouz PJ, Lahey BB. External validity of a hierarchical dimensional model of child and adolescent psychopathology: Tests using confirmatory factor analyses and multivariate behavior genetic analyses. Journal of abnormal psychology. 2016; 125(8):13.

33. Schaefer JD, Caspi A, Belsky DW, Harrington H, Houts R, Horwood LJ, et al. Enduring mental health: Prevalence and prediction. Journal of abnormal psychology. 2017; 126(2):212. https://doi.org/10.1037/abn0000232 PMID: 27929304

34. Krueger RF, Kotov R, Watson D, Forbes M, Eaton N, Ruggero C, et al. Progress in achieving quantitative classification of psychopathology. World Psychiatry. in press.

35. Ravens-Sieberer U, Gosch A, Rajmil L, Erhart M, Bruil J, Power M, et al. The KIDSCREEN-52 Quality of Life Measure for Children and Adolescents: Psychometric Results from a Cross-Cultural Survey in 13 European Countries. Value in Health. 2008; 11(4):645–58. https://doi.org/10.1111/j.1524-4733.2007.00291.x PMID: 18179669

36. Clarke A, Friede T, Putz R, Ashdown J, Martin S, Blake A, et al. Warwick-Edinburgh Mental Well-being Scale (WEMWBS): Validated for teenage school students in England and Scotland. A mixed methods assessment. BMC Public Health. 2011; 11(1):487.

37. Spinhoven P, Elzinga BM, Giltay E, Penninx BWJH. Anxious or Depressed and Still Happy? PLOS ONE. 2015; 10(10):e0139912. https://doi.org/10.1371/journal.pone.0139912 PMID: 26461261

38. Keyes CLM. Mental Illness and/or Mental Health? Investigating Axioms of the Complete State Model of Health. Journal of Consulting and Clinical Psychology. 2005; 73(3):539–48. https://doi.org/10.1037/0022-006X.73.3.539 PMID: 15982151

39. Westerhof GJ, Keyes CLM. Mental Illness and Mental Health: The Two Continua Model Across the Lifespan. Journal of Adult Development. 2010; 17(2):110–9. https://doi.org/10.1007/s10804-009-9082-y PMID: 20502508

40. Böhnke JR, Croudace TJ. Calibrating well-being, quality of life and common mental disorder items: psychometric epidemiology in public mental health research. The British Journal of Psychiatry. 2016; 209(2):162–8. https://doi.org/10.1192/bjp.bp.115.165530 PMID: 26635327

41. St Clair MC, Neufeld S, Jones PB, Fonagy P, Bullmore ET, Dolan RJ, et al. Characterising the latent structure and organisation of self-reported thoughts, feelings and behaviours in adolescents and young adults. PLOS ONE. 2017; 12(4):e0175381. https://doi.org/10.1371/journal.pone.0175381 PMID: 28403164

42. Diener E. Subjective well-being. The science of happiness and a proposal for a national index. The American psychologist. 2000; 55(1):34–43. PMID: 11392863

43. Ryan RM, Deci EL. On Happiness and Human Potentials: A Review of Research on Hedonic and Eudaimonic Well-Being. Annual Review of Psychology. 2001; 52(1):141–66.

44. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). Washington, UNITED STATES: American Psychiatric Publishing; 2013.

45. Vanhoutte B. The Multidimensional Structure of Subjective Well-Being In Later Life. Journal of Population Ageing. 2014; 7(1):1–20. https://doi.org/10.1007/s12062-014-9092-9 PMID: 25089162

46. Achenbach TM, Edelbrock CS. Psychopathology of Childhood. Annual Review of Psychology. 1984; 35(1):227–56.

47. Goodman A, Lamping DL, Ploubidis GB. When to Use Broader Internalising and Externalising Subscales Instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British Parents, Teachers and Children. Journal of Abnormal Child Psychology. 2010; 38(8):1179–91. https://doi.org/10.1007/s10802-010-9434-x PMID: 20623175

48. Gutman L, Joshi H, Parsonage M, Schoon I. Children of the new century. Mental health findings from the Millennium Cohort Study London: Centre for Mental Health2015.

49. Fink E, Patalay P, Sharpe H, Holley S, Deighton J, Wolpert M. Mental Health Difficulties in Early Adolescence: A Comparison of Two Cross-Sectional Studies in England From 2009 to 2014. Journal of Adolescent Health. 2015; 56(5):502–7. https://doi.org/10.1016/j.jadohealth.2015.01.023 PMID: 25907650

50. Fitzsimons E, Goodman A, Kelly E, Smith JP. Poverty dynamics and parental mental health: Determinants of childhood mental health in the UK. Social Science & Medicine. 2017; 175:43–51.

51. Kinderman P, Schwannauer M, Pontin E, Tai S. Psychological Processes Mediate the Impact of Familial Risk, Social Circumstances and Life Events on Mental Health. PLOS ONE. 2013; 8(10):e76564. https://doi.org/10.1371/journal.pone.0076564 PMID: 24146890

52. Bamber D, Tamplin A, Park RJ, Kyte ZA, Goodyer IM. Development of a Short Leyton Obsessional Inventory for Children and Adolescents. Journal of the American Academy of Child & Adolescent Psychiatry. 2002; 41(10):1246–52.

53. Raine A. The SPQ: a scale for the assessment of schizotypal personality based on DSM-III-R criteria. Schizophrenia bulletin. 1991; 17(4):555. PMID: 1805349

54. Reynolds CR, Richmond BO. What i think and feel: A revised measure of children's manifest anxiety. Journal of Abnormal Child Psychology. 1978; 6(2):271–80. PMID: 670592

55. Horwood J, Salvi G, Thomas K, Duffy L, Gunnell D, Hollis C, et al. IQ and non-clinical psychotic symptoms in 12-year-olds: results from the ALSPAC birth cohort. British Journal of Psychiatry. 2008; 193 (3):185–91. https://doi.org/10.1192/bjp.bp.108.051904 PMID: 18757973

56. Eid M, Geiser C, Koch T, Heene M. Anomalous results in G-factor models: Explanations and alternatives. Psychological Methods. 2017; 22(3):541–62. https://doi.org/10.1037/met0000083 PMID: 27732052

57. van Bork R, Epskamp S, Rhemtulla M, Borsboom D, van der Maas HLJ. What is the p-factor of psychopathology? Some risks of general factor modeling. Theory & Psychology. 2017; 27(6):759–73.

58. Murray AL, Johnson W. The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. Intelligence. 2013; 41(5):407–22.

59. Reise SP, Kim DS, Mansolf M, Widaman KF. Is the Bifactor Model a Better Model or Is It Just Better at Modeling Implausible Responses? Application of Iteratively Reweighted Least Squares to the Rosenberg Self-Esteem Scale. Multivariate Behavioral Research. 2016; 51(6):818–38. https://doi.org/10.1080/00273171.2016.1243461 PMID: 27834509

60. Bonifay W, Lane SP, Reise SP. Three Concerns With Applying a Bifactor Model as a Structure of Psychopathology. Clinical Psychological Science. 2017; 5(1):184–6.

61. Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling:A Bifactor Perspective. Educational and Psychological Measurement. 2013; 73(1):5–26.

62. Rodriguez A, Reise SP, Haviland MG. Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. Journal of Personality Assessment. 2016; 98(3):223–37. https://doi.org/10.1080/00223891.2015.1089249 PMID: 26514921

63. Ten Berge JMF, Sočan G. The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. Psychometrika. 2004; 69(4):613–25.

64. Anna Freud Centre. HeadStart Pilot Evaluation [internet]. London: Anna Freud Centre; n.d. [Available from: https://www.annafreud.org/what-we-do/research-policy/research-themes/improving-and-evaluating-services/headstart-pilot-evaluation/.

65. Department for Education. Special educational needs in England: January 2015. 2015.

66. Department for Education. Schools, pupils and their characteristics: January 2015. 2015.

67. Panayiotou M, Humphrey N. Mental health difficulties and academic attainment: Evidence for gender-specific developmental cascades in middle childhood. Development and Psychopathology. 2017:1–16.

68. Humphrey N, Wigelsworth M. Making the case for universal school-based mental health screening. Emotional and Behavioural Difficulties. 2016; 21(1):22–42.

69. Deighton J, Tymms P, Vostanis P, Belsky J, Fonagy P, Brown A, et al. The Development of a School-Based Measure of Child Mental Health. Journal of Psychoeducational Assessment. 2013; 31(3):247–57. https://doi.org/10.1177/0734282912465570 PMID: 25076806

70. Goodman R. The Strengths and Difficulties Questionnaire: A Research Note. Journal of Child Psychology and Psychiatry. 1997; 38(5):581–6. PMID: 9255702

71. Patalay P, Deighton J, Fonagy P, Vostanis P, Wolpert M. Clinical validity of the Me and My School questionnaire: a self-report mental health measure for children and adolescents. Child and Adolescent Psychiatry and Mental Health. 2014; 8(1):17.

72. Duncan B, Sparks J, Miller S, Bohanske R, Claud D. Giving Youth a Voice: A Preliminary Study of the Reliability and Validity of a Brief Outcome Measure for Children, Adolescents, and Caretakers. Journal of Brief Therapy. 2006; 5(2):71–88.

73. Gorard S. A cautionary note on measuring the pupil premium attainment gap in England. British journal of education, society and behavioural science. 2016; 14(2):1–8.

74. Brown TA. Confirmatory factor analysis for applied research: Guilford Publications; 2015.

75. Li C-H. Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. Behavior Research Methods. 2016; 48(3):936–49. https://doi.org/10.3758/s13428-015-0619-7 PMID: 26174714

76. Hox J, Maas C, Brinkhuis M. The effect of estimation method and sample size in multilevel structural equation modeling. Statistica Neerlandica. 2010; 64(2):157–70.

77. Mutheén BO, Mutheén LK, Asparouhov T. Estimator choices with categorical outcomes. Mplus and Mplus2015.

78. Lt Hu, 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. Raykov T, Marcoulides GA. Introduction to psychometric theory: Routledge; 2011.

80. Raykov T, Marcoulides GA, Patelis T. The Importance of the Assumption of Uncorrelated Errors in Psychometric Theory. Educational and Psychological Measurement. 2015; 75(4):634–47. https://doi.org/10.1177/0013164414548217 PMID: 29795836

81. Cramer AOJ, Sluis S, Noordhof A, Wichers M, Geschwind N, Aggen SH, et al. Measurable Like Temperature or Mereological Like Flocking? On the Nature of Personality Traits. European Journal of Personality. 2012; 26(4):451–9.

82. Mutheén LK, Mutheén BO. Mplus User's Guide. Eighth Edition. Los Angeles, CA: Mutheén & Mutheén; 1998–2017.

83. Reise SP. The Rediscovery of Bifactor Measurement Models. Multivariate Behavioral Research. 2012; 47(5):667–96. https://doi.org/10.1080/00273171.2012.715555 PMID: 24049214

84. Bowen NK, Masa RD. Conducting Measurement Invariance Tests with Ordinal Data: A Guide for Social Work Researchers. Journal of the Society for Social Work and Research. 2015; 6(2):229–49.

85. Asparouhov T, Mutheén BO. Nesting and Equivalence Testing in Mplus [internet]. 2018 [1–17]. Available from: http://www.statmodel.com/download/NET.pdf.

86. Mutheén LK. Negative Residual Variance [internet]. 2007 [Available from: http://www.statmodel.com/discussion/messages/9/572.html?1500932974.

87. Byrne BM, Shavelson RJ, Muthén B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin. 1989; 105(3):456–66.

88. Millsap RE, Kwok OM. Evaluating the impact of partial factorial invariance on selection in two populations. Psychol Methods. 2004; 9(1):93–115. https://doi.org/10.1037/1082-989X.9.1.93 PMID: 15053721

89. Sass DA. Testing Measurement Invariance and Comparing Latent Factor Means Within a Confirmatory Factor Analysis Framework. Journal of Psychoeducational Assessment. 2011; 29(4):347–63.

90. Cheung GW, Rensvold RB. Testing Factorial Invariance across Groups: A Reconceptualization and Proposed New Method. Journal of Management. 1999; 25(1):1–27.

91. Kenny DA, McCoach DB. Effect of the Number of Variables on Measures of Fit in Structural Equation Modeling. Structural Equation Modeling: A Multidisciplinary Journal. 2003; 10(3):333–51.

92. Lilienfeld SO. Comorbidity Between and Within Childhood Externalizing and Internalizing Disorders: Reflections and Directions. Journal of Abnormal Child Psychology. 2003; 31(3):285–91. PMID: 12774861

93. Moilanen KL, Shaw DS, Maxwell KL. Developmental cascades: Externalizing, internalizing, and academic competence from middle childhood to early adolescence. Development and Psychopathology. 2010; 22(3):635–53. https://doi.org/10.1017/S0954579410000337 PMID: 20576184

94. Borsboom D, Cramer AOJ. Network Analysis: An Integrative Approach to the Structure of Psychopathology. Annual Review of Clinical Psychology. 2013; 9(1):91–121.

95. Marsh JK, De Los Reyes A. Explaining away disorder: The influence of context on impressions of mental health symptoms. Clinical Psychological Science. 2018; 6(2): 189–202.

96. De Los Reyes A, Augenstein TM, Wang M, Thomas SA, Drabick DAG, Burgers DE, Rabinowitz J. The validity of the multi-informant approach to assessing child and adolescent mental health. Psychological Bulletin. 2015; 141(4):858–900. https://doi.org/10.1037/a0038498 PMID: 25915035

97. Patalay P, Fitzsimons E. Mental ill-health among children of the new century: Trends across childhood with a focus on age 14. London: Centre for Longitudinal Studies; 2017.

98. Patalay P, Fitzsimons E. Development and predictors of mental ill-health and wellbeing from childhood to adolescence. Social Psychiatry and Psychiatric Epidemiology. 2018; 53(12):1311–1323. https://doi.org/10.1007/s00127-018-1604-0 PMID: 30259056

99. Kline R. Principles and practice of structural equation modeling Fourth Edition. New York: The Guilford Press; 2015.

100. Ilie S, Sutherland A, Vignoles A. Revisiting free school meal eligibility as a proxy for pupil socio-economic deprivation. British Educational Research Journal. 2017; 43(2):253–74.

Implications of Paper 1 for Policy Makers and/or Educators and Teachers

Previous work has tended to emphasize the following policy implications: positive and negative mental health should be used together in assessment; interventions could have differential effects for each of symptoms and wellbeing; wellbeing measurement should be integrated into public systems to aid prevention and understand strengths as well as weaknesses (Iasiello & Agteren, 2020). The findings of Paper 1 both support and challenge these ideas. First, for the question of joint assessment and integration into health/education systems, the finding of particular similarity between internalizing symptoms and wellbeing suggests positive wellbeing may be particularly useful for early identification of internalizing symptoms.

However, this similarity (equivalent to that between the two symptom domains) raises the issue of how these outcomes should be compared. Given that symptom domains are often grouped together in dual-factor research, the distinction between wellbeing and symptoms seems hard to interpret in light of Paper 1's findings: If internalizing symptoms are as similar to wellbeing as to externalizing symptoms, is it justified to argue that positive mental health is the separate construct? On the basis of Paper 1, it is therefore not recommended to assume that symptoms (including a range of domains) are inherently homogenous, nor that wellbeing is similarly inherently distinct. This has implications for assessment but also for intervention, particularly judging response to this. In either assessment or intervention response the question of differences between symptoms and wellbeing should therefore be considered in light of conceptual similarities and differences (see also subsequent papers). Findings nevertheless tentatively support the idea that positive approaches could be useful for early identification of problems to aid prevention efforts, given the strong relationship between wellbeing and internalizing symptoms. Teachers and policy makers should therefore consider wellbeing measures as potentially useful tools for early identification, and employ some scepticism when considering interventions/measures that claim to *only* consider one aspect of positive or negative mental health.

References

Iasiello, M., & Agteren, J. v. (2020). *Mental health and/or mental illness: A scoping review of the evidence and implications of the dual-continua model of mental health*. Exeley. https://doi.org/10.3316/informit.261420605378998

Paper 2: Internalizing Symptoms, Well-being, and Correlates in Adolescence: A Multiverse Exploration Via Cross-Lagged Panel Network Models

Status: published, gold open access

Black, L., Panayiotou, M., & Humphrey, N. (2021). Internalizing symptoms, well-being, and correlates in adolescence: A multiverse exploration via cross-lagged panel network models. Development and Psychopathology, 1-15. doi:10.1017/S0954579421000225

Supplementary material can be found at https://osf.io/rxv5q/

Development and Psychopathology (2021), 1–15
doi:10.1017/S0954579421000225

CAMBRIDGE
UNIVERSITY PRESS

Regular Article

Internalizing symptoms, well-being, and correlates in adolescence: A multiverse exploration via cross-lagged panel network models

Louise Black, Margarita Panayiotou and Neil Humphrey
Manchester Institute of Education, University of Manchester, Manchester, UK

Abstract

Internalizing symptoms are the most prevalent mental health problem in adolescents, with sharp increases seen, particularly for girls, and evidence that young people today report more problems than previous generations. It is therefore critical to measure and monitor these states on a large scale and consider correlates. We used novel panel network methodology to explore relationships between internalizing symptoms, well-being, and inter/intrapersonal indicators. A multiverse design was used with 32 conditions to consider the stability of results across arbitrary researcher decisions in a large community sample over three years ($N = 15{,}843$, aged 11–12 at Time 1). Networks were consistently similar for girls and boys. Stable trait-like effects within anxiety, attentional, and social indicators were found. Within-person networks were densely connected and suggested mental health and inter/intrapersonal correlates related to one another in similar complex ways. The multiverse design suggested the particular operationalization of items can substantially influence conclusions. Nevertheless, indicators such as thinking clearly, unhappiness, dealing with stress, and worry showed more consistent centrality, suggesting these indicators may play particularly important roles in the development of mental health in adolescence.

Keywords: adolescence, mental health, network psychometrics, well-being

(Received 1 September 2020; revised 28 January 2021; accepted 9 March 2021)

Adolescence is recognized as a key developmental phase characterized by rapid physical, social, and psychological change (Dahl, Allen, Wilbrecht, & Suleiman, 2018; Patton et al., 2016; Sawyer, Azzopardi, Wickremarathne, & Patton, 2018). The majority of lifetime disorders also show first onset in the teenage years (Jones, 2013). Early adolescence is likely particularly important to understanding what sets in motion changes in mental health. For instance, key gender differences emerge, and contextual factors such as puberty and school transition are in process (Patalay & Fitzsimons, 2017, 2018). Evidence of the correlates of mental health in this age could therefore be key to improving identification, intervention, and prevention (Patalay & Fitzsimons, 2018). However, based on available evidence, methodological challenges make it difficult to determine which indicators are particularly important (see sections outlining analytical considerations below). This study therefore makes use of a new panel network model (Epskamp, 2020b) to consider indicator-level interactions (pairwise causal associations that are often bidirectional; Epskamp, Rhemtulla, & Borsboom, 2017) between internalizing symptoms, well-being, and inter/intrapersonal indicators.

Sharp increases in internalizing problems, particularly for girls, make up much of the mental health difficulties faced by adolescents (Rapee et al., 2019), and evidence suggests levels of these problems are increasing over time (Collishaw, 2015; NHS Digital, 2018). Furthermore, internalizing problems are also highly comorbid with other disorders (Carrellas, Biederman, & Uchida, 2017; Merikangas et al., 2010; NHS Digital, 2018; Wolff & Ollendick, 2006) and substantially correlated with other symptoms (Black, Panayiotou, & Humphrey, 2019; Patalay et al., 2015), making them an important focus for inquiry. Large numbers of adolescents also experience subthreshold internalizing symptoms. For instance, up to 12% of 11–14-year-olds experience subthreshold levels of depression (Bertha & Balázs, 2013). Here we consider the key theoretical and analytical considerations in the robust study of internalizing problems.

Theoretical considerations

The role of well-being

Further insight into internalizing problems and those at risk might be afforded by also measuring well-being (Bartels, Cacioppo, van Beijsterveldt, & Boomsma, 2013). This reflects the World Health Organization's longstanding definition that mental health should not consist only of the absence of symptoms (WHO, 1946). This broader conceptualization is also likely to be more useful in nonclinical samples, since positive mental health can capture greater variability (Alexander, Salum, Swanson, & Milham, 2020). Well-being is also closely related to internalizing symptoms, statistically and conceptually (Black, Panayiotou, & Humphrey, 2020b), showing substantial correlations for total

Author for Correspondence: Louise Black, Manchester Institute of Education, University of Manchester, Manchester, M13 9PL; E-mail: louise.black@manchester.ac.uk.

Cite this article: Black L, Panayiotou M, Humphrey N (2021). Internalizing symptoms, well-being, and correlates in adolescence: A multiverse exploration via cross-lagged panel network models. *Development and Psychopathology* 1–15. https://doi.org/10.1017/S0954579421000225

scores and latent constructs (.41–.68; Antaramian, Huebner, Hills, & Valois, 2010; Black et al., 2019; Suldo, Thalji, & Ferron, 2011). Correlations around this level suggest that constructs are substantially related while each still contributes distinct information.

Furthermore, given that self-report adolescent mental health problem data can be error-prone, it can be argued that well-being might be used to strengthen measurement. Specifically, substantial measurement error in adolescent mental health problems is suggested by low inter-rater associations and varying approaches to classification, and there is no clear criterion against which such measures can be validated (Wolpert & Rutter, 2018). Commonly used symptom measures are typically old and/or based on limited psychometric investigation (Bentley, Hartley, & Bucci, 2019; Black, Mansfield, & Panayiotou, 2020a; Dedrick, Greenbaum, Friedman, Wetherington, & Knoff, 1997; Goodman, 2001), whereas newer well-being measures that followed modern and rigorous item-development and validation standards (e.g., Ravens-Sieberer et al., 2005; Stewart-Brown et al., 2009), may complement symptom data and improve measurement accuracy. Routine adoption of such measures is also empirically justified since well-being seems to relate at a similar level to different domains of psychopathology as these relate to one another (e.g., Black et al., 2019). Since these psychopathology domains have been amalgamated into composites (e.g., Patalay & Fitzsimons, 2016), and there is conceptual and statistical similarity at the indicator level for internalizing symptoms and well-being (Black et al., 2020b), using well-being measures to capture additional information can be a useful approach.

Intra and interpersonal correlates of internalizing symptoms in adolescence

There is a substantial body of literature covering the developmental risk and promotive correlates of mental health in adolescents (for reviews see for example, Evans, Li, & Whipple, 2013; Fritz, de Graaff, Caisley, van Harmelen, & Wilkinson, 2018; Masten & Barnes, 2018). Moreover, there is theoretical consensus that systems models are appropriate (Bronfenbrenner, 2005; Evans et al., 2013; Masten & Barnes, 2018), namely considering factors from across personal (e.g., problem solving), family (e.g., secure attachment), and wider environments (e.g., school connectedness), and key correlates have consistently been identified across samples and methods (Masten & Barnes, 2018). It is important to capture these multiple systems since effects can cascade from one level to the other such that the interaction between mental health and environments is inherently complex (Masten & Cicchetti, 2010). Consistent with other literature considering the dynamic interplay of correlates and mental health, we focus on malleable (i.e., intra and interpersonal factors) rather than biological or socioeconomic variables (Fritz et al., 2019).

For internalizing problems in adolescence specifically, it is thought that social factors and emotional regulation are particularly key factors (Rapee et al., 2019), suggesting these should be particularly studied in the development of internalizing symptoms. The sudden physical, psychological, and social changes experienced in adolescence might affect expectations and views of young people, and these changes likely in turn impact internalizing symptoms (Rapee et al., 2019). Perceived home, peer, and school support are therefore likely important correlates. More generally, emotion regulation can be impacted by difficult home environments (e.g., maternal depression or parental conflict), and resulting difficulties managing emotions pose significant risk for internalizing problems (Thompson, 2019).

Gender differences

Inclusion of such correlates also facilitates consideration of a key issue for internalizing symptoms in adolescence, namely that these disproportionately affect girls (Merikangas et al., 2010; NHS Digital, 2018), and that this is increasingly the case (Bor, Dean, Najman, & Hayatbakhsh, 2014; Collishaw, 2015). A key theme in the theoretical literature is whether girls and boys experience quantitatively or qualitatively different risk factors (Hyde, Mezulis, & Abramson, 2008). Indicator-level analysis of internalizing symptoms, well-being, and relevant malleable correlates over time may therefore shed light on this question. For instance, it may be that previous construct-level analyses have made differences difficult to pin-point with variation occurring (qualitatively) at the indicator level. Alternatively, if a common network structure that varies in edge strength is found, quantitative differences may explain prevalence findings.

Analytical considerations

Within- and between-person effects

Longitudinal data consist of both variation within individuals (over time), and variation between individuals (Curran & Bauer, 2011). In panel data, people are nested in time, much as in multilevel data, for instance, children are nested in schools. This allows for the consideration of how variables influence one another within people on average over time, taking account of stable (or trait-like) individual differences. In the estimation of the cross-lagged panel network, estimated stable means, and deviations from these over time allow for a network of trait-like effects over time, a longitudinal network of malleable effects over time, and a contemporaneous network of (undirected) state-like effects that happen within the lag considered (in our case more quickly than once a year). For example, adolescents' general tendencies to report anxiety might be related to their general tendencies to report perceived social support. This trait-like effect therefore needs to be controlled for when considering the direction and strength of the temporal association between anxiety and social support.

This kind of disaggregation has led to new findings in construct-level panel models. For instance, while bidirectional relationships have been observed for internalizing and externalizing symptoms, only the latter predicted the former when disaggregated effects were considered (Flouri et al., 2019; Oh et al., 2020). Similar findings have been observed for adolescent depression and self-esteem (Masselink et al., 2018). There is also early evidence in younger children that correlates at different ecological levels can interact reciprocally at the within-person level (after accounting for between-person effects). Kaufman, Kretschmer, Huitsing, and Veenstra (2020) found evidence of such effects for internalizing symptoms, parenting, and bullying.

Thus, to understand how temporal effects between psychological variables occur for the average individual, analysis of within-person effects, accounting for between-person differences, is needed. For instance, we might consider whether change in internalizing problems is predicted by bullying. Without disaggregated analysis, and assuming other requirements for causal inference are met (Rohrer, 2018), we cannot be sure that those experiencing symptoms are not in fact also those commonly targeted by bullies (a between-person effect). Crucially, while it is well established that disaggregation of within and between-person effects is needed for accurate inferences to be made, it is still common-place to assume within-person processes from analyses representing a blend of within and between variance (Hamaker, Kuiper, & Grasman, 2015).

Network analysis

While the studies cited above have modeled within and between-person effects separately, they have relied on total scores and latent factors which treat individual symptoms as indicators of a given mental state. While this approach can be statistically equivalent (Fried, 2020), we argue it is theoretically problematic, given the absence of external evidence for disorders, the likelihood that mental health states are contributed to by a constellation of biological and environmental factors, and the fact many disorders share indicators (Borsboom, Cramer, & Kalis, 2018; Borsboom, Cramer, Schmittmann, Epskamp, & Waldorp, 2011). Network approaches might better capture the nuance and complexity likely to be present in adolescent mental health (Kalisch et al., 2019). These offer the opportunity to consider individual indicators and correlates as outcomes and predictors while accounting for all other indicators in the model (Epskamp, 2020b; Kalisch et al., 2019). For example, from the network perspective we can consider the unique association of bullying and worry, and in longitudinal networks we can also track direction. Mental states and their correlates can therefore be represented as dynamic with interactive indicators, such that, for instance, bullying leads to worry, which in turn leads to somatic symptoms, which in turn leads to unhappiness. The modeling of indicators, and not latent variables, within network models is arguably particularly appropriate for internalizing symptoms in adolescence since evidence suggests a lack of clear clustering into theoretical disorders (e.g., depression, anxiety) for this domain (McElroy & Patalay, 2019; McElroy, Fearon, Belsky, Fonagy, & Patalay, 2018).

Item-level differences in reporting have also been found in young adolescents for internalizing and well-being (Black et al., 2019). Similarly, analysis of adult samples suggest indicator-level analysis could be important to understanding gender differences. Fried, Nesse, Zivin, Guille, and Sen (2014) found men reported more suicidal ideation and psychomotor symptoms of depression in response to stress, while women reported more fatigue, appetite, and sleep problems. Within-person analysis at the indicator-level could therefore be key to improving understanding of the development of internalizing symptoms, including gender differences.

The current study

The current study aimed to explore indicator-level within and between-person associations for internalizing symptoms, well-being, and inter/intrapersonal correlates via novel panel network models (Epskamp, 2020b). A conceptual demonstration of the panel network model, is shown in Figure 1. This diagram is simplified to aid interpretation and therefore shows parameters for only two indicators, while in the current study 22 are included. The existence of large panel studies represents an opportunity to consider longitudinal indicator-level associations in rich datasets, in which within-person effects can be modeled (Curran & Bauer, 2011). We therefore conducted secondary analysis of a dataset designed to explore and test new ways to improve mental health and well-being of young people aged 10–16. The current study was based on existing data which we were familiar with the HeadStart (HS) evaluation (Deighton et al., 2019). Therefore, a multiverse approach in which multiple combinations of possible reasonable decisions are analyzed in parallel, was used to avoid researcher degrees of freedom obscuring results (Simmons, Nelson, & Simonsohn, 2011; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; Weston, Ritchie, Rohrer, & Przybylski, 2019).

Given gender differences for internalizing symptoms and personal and social resources (Rapee et al., 2019), we expected that associations in within and between models would be noninvariant across girls and boys. We hypothesized that irrespective of gender (a) social problems (e.g., being bullied) would show positive associations with internalizing symptoms and negative with well-being; (b) well-being and symptoms would be negatively associated; (c) intrapersonal factors (e.g., the ability to handle stress) would be negatively associated with symptoms and positively with well-being; (d) social support would be negatively associated with symptoms and positively associated with well-being. Given the lack of studies analyzing disaggregated models, we were unable to specify which effects would be observed at within or between-person levels. Finally, we explored which indicators were the most influential and the most predicted.

Method

Background and procedure

We undertook secondary analysis of data from three annual waves (2017–2019) collected from a longitudinal cohort study. The project from which data were drawn aims to explore and test new ways to improve mental health and well-being of young people aged 10–16 and prevent serious mental health issues from developing.

Ethical approval was granted by the UCL ethics committe (reference: 8097/003), and opt-out parental consent was given for adolescents to complete secure online surveys during the school day. Teachers read out an information sheet which emphasized pupils' confidentiality and right to withdraw. Socio-demographic data were drawn from the National Pupil Database.

Participants

Data were collected from 15,859 pupils in year seven (age 11–12) at Time 1, from 118 secondary schools in England (52.7% female). Given the focus of the project, the sample was not drawn to be representative: 35.4% had ever been eligible for free school meals at Time 1 compared to the national figure of 28.5% eligible in the previous six years (Department for Education, 2017a); 12.0% had special educational needs (national figure = 14.4%; Department for Education, 2017c); in terms of ethnicity, 74.2% were white (national figure = 75.2%), 9.3% were Asian (national figure = 10.7%), 5.7% were of Black origin (national figure = 5.6%), 4.0% were of mixed origin (national figure = 5.0%), .2% were Chinese (national figure = .45), while 1.6% were classified as any other ethnic group (national figure = 1.75), and 1.5% were unclassified (national figure = 1.5%; Department for Education, 2017b). Of this total sample, 16 were removed from the current study since they had missing data for all items included for analysis.

Item selection

The conceptual domains explored in the current study (based on the literature reviewed above and indicators available in the dataset at each time point) were: internalizing symptoms (including

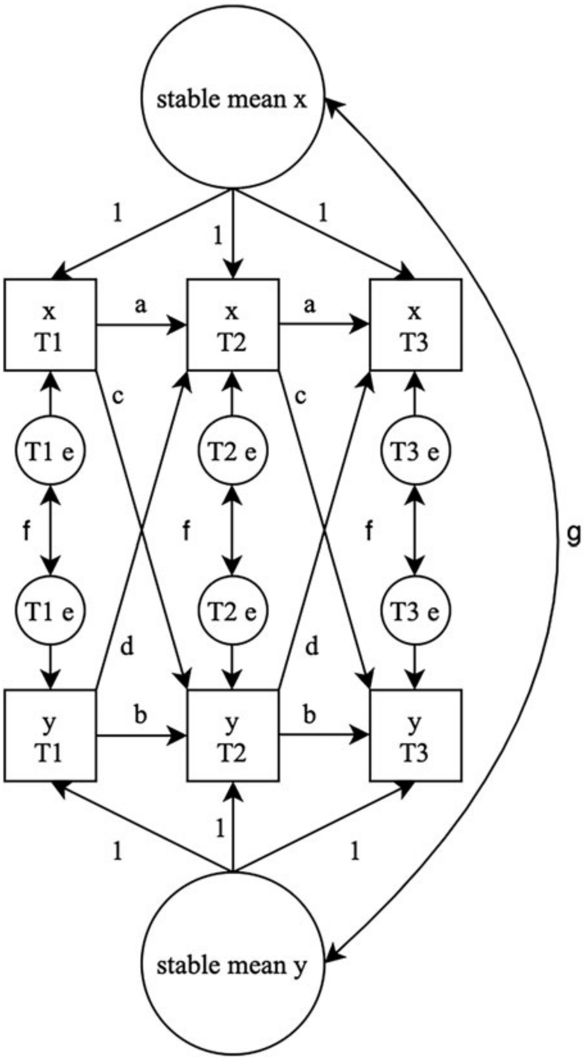


Figure 1. Conceptual diagram of a panel network model for two indicators, x and y, at three time points, T1–T3. Paths a–d represent average within-person directed partial correlations, including autocorrelations (temporal network). Paths marked f represent within-person partial correlations within lags (contemporaneous networks), with e representing the residual for each indicator after accounting for temporal effects. Path g represents between-person partial correlations for stable trait-like effects (between network).

attentional symptoms and social withdrawal; American Psychiatric Association, 2013; WHO, 2018), well-being, home, school and peer support, and intrapersonal factors such as managing stress. The choice of indicators was restricted, given that the software used for the panel network analysis currently cannot handle more than around 30 (Epskamp, 2020b) and it is not appropriate to indiscriminately include highly similar indicators in networks (Fried & Cramer, 2017; Rhemtulla, Cramer, van Bork, & Williams, 2018). Items were therefore selected from those available in the dataset according to the following criteria: (a) conceptual domain, (b) item simplicity, given issues highlighted in this area (Black et al., 2020a), (c) descriptive and factor model statistics.

The final list of items is shown in Table 1 alongside descriptive statistics (full item wording is available in the supplementary material, S1). Items were drawn from the Strengths and Difficulties Questionnaire (SDQ, Goodman, Meltzer, & Bailey, 1998), Short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS, Stewart-Brown et al., 2009), Student Resilience Survey (SRS, Lereya et al., 2016), Trait Emotional Intelligence Questionnaire-Adolescent Short Form (TEIQUE-ASF, Petrides & Furnham, 2009), and four-item perceived stress scale (Demkowicz, Panayiotou, Ashworth, Humphrey, & Deighton, 2019).

Multiverse approach

In order to increase transparency, sensitivity analyses of many possible analytical decisions were conducted (Steegen et al., 2016; Weston et al., 2019). In line with multiple specification approaches (Simonsohn, Simmons, & Nelson, 2020; Steegen et al., 2016), variation in decisions was limited to those we considered likely to provide valid insight. In addition, given the novelty and computationally demanding nature of the analyses presented here, we also limited conditions based on feasibility. For instance, given that valid inferences can be drawn across a wide range of search algorithms at large sample sizes (Epskamp, 2020b), we used only two such robust, but relatively computationally light, procedures.[1]

Two aspects were identified as vulnerable to researcher degrees of freedom. First, the choice of items was in some cases arbitrary such that there were items from more than one scale that considered the same relevant experience. Second, the novelty of the method means that which estimation algorithm is most appropriate has not been clearly established. In such instances, multiverse approaches are recommended (Epskamp, 2019). This resulted in 16 possible datasets (based on varying two possible item operationalizations for four items: *distracted, mind*, *optimism*, *problem*, see Table 1) × two search algorithms, meaning that models for 32 conditions were estimated. For more details on the choice of items see the supplementary material (S1). Only full information maximum likelihood (FIML) estimation was selected since data cannot yet be treated as ordinal in the panel network model and no robust adjustments are available. Two equally robust pruning methods were considered: alpha at .01, and this plus stepwise modification based on the Bayesian information criterion (BIC).

We also kept the number of indicators in each model constant since we wanted to avoid overfitting by including multiple indicators of the same experience (e.g., two peer support items), and since networks are not directly comparable with varying numbers of nodes (Costantini et al., 2019). Missing data were retained for all conditions given FIML estimation, and since analysis was at the item level, and data were ordinal, outliers were not considered.

The resulting design allowed us to assess the stability of the most influential and predicted indicators and gender invariance across these decisions. Fit was assumed to be good across conditions given the data-driven approach, and was not used to compare conditions. Since our analysis was exploratory, testing multiple contingent effects, we approached the results of our sensitivity analyses descriptively in line with Steegen et al. (2016). We therefore present how fit, strength, and gender invariance varied across analyses.

Analysis

Code for all analyses, and simulated data for the purpose of running code, is available in the supplementary material (S2–S4). In

[1] We found it was not possible to run the modelsearch function with item-level panel analysis in *psychonetrics*.

Table 1. Node names, item wording and descriptive statistics over time

Node name/abbreviated wording	Measure	Mean (*SD*)	Skew
Bully/others bully me	SDQ	.31–.42 (.58–.66)	1.3–1.74
Close/feeling close to others	SWEMWBS	2.33–2.41 (1.10–1.15)	.51–.60
Distracted A/easily distracted	SDQ	1.01– 1.07 (.74–.76)	–.01–.11
Distracted B/my attention is good	SDQ	.76–.86 (.64)	.13–.25
Feelings/hard to control my feelings	TEIQUE-ASF	3.17–3.29 (1.9–2.0)	.47–.58
Help/when in need I find someone	SRS	2.38–2.75 (1.31–1.34)	.23–.58
Home/at home there is an adult who listens	SRS	1.76–1.87 (1.02–1.09)	1.14–1.31
Mind A/able to make up my own mind	SWEMWBS	2.10–2.25 (1.05)	.63–.79
Mind B/change my mind often	TEIQUE-ASF	3.17–.3.29 (1.90–2.0)	.47–.58
Nervous/nervous in new situations…easily lose confidence	SDQ	.99–1.09 (.76)	–.15–.02
Optimism A/optimistic about the future	SWEMWBS	2.64–2.69 (1.06–1.13)	.26–.33
Optimism B/things were going your way	PSS-4	1.93–1.98 (1.05–1.06)	.12–.14
Peer/there are students at school who make you feel better	SRS	1.93–2.02 (1.15–1.16)	.97–1.14
Problem A/dealing with problems well	SWEMWBS	2.64–2.74 (1.11–1.16)	.27–.32
Problem B/ability to handle personal problems	PSS-4	1.68–1.76 (1.13–1.19)	.23–.30
Relaxed/feeling relaxed	SWEMWBS	2.72–2.90 (1.11–1.12)	.03–.18
Restless/restless…cannot stay still	SDQ	1.10–1.12 (.73–.74)	–.15–.19
Scared/many fears…scared	SDQ	.61–.65 (.71–.72)	.65–.72
School/at school an adult really cares about me	SRS	2.46–2.74 (1.24–1.28)	.21–.45
Somatic/headaches, stomach-aches or sickness	SDQ	.74–.78 (.73–.76)	.40–.45
Stress/able to deal with stress	TEIQUE-ASF	3.72–3.85 (1.97–2.10)	.06–.14
Think/thinking clearly	SWEMWBS	2.47–2.70 (1.08)	.28–.41
Unhappy/unhappy, downhearted or tearful	SDQ	.53–.62 (.68–.71)	.70–.90
Useful/feeling useful	SWEMWBS	2.74–2.86 (1.04–1.07)	.11–.26
Withdrawn/usually on my own	SDQ	.43–.45 (.65–.67)	1.18–1.23
Worry/worry a lot	SDQ	.94–1.04 (.77–.79)	–.07–.11

Note. SDQ = strengths and difficulties questionnaire; SWEMWBS = short Warwick-Edinburgh mental well-being scale; TEIQUE-ASF = trait emotional intelligence questionnaire- adolescent short form; PSS-4 = 4-item perceived stress scale; SRS = student resilience survey.

order to count relationships counter to our hypotheses across conditions, all indicators were coded to have a positive manifold (e.g., well-being indicators were reversed with respect to symptoms). The first stage of the main analysis (for each condition) was to estimate a panel network model for each whole sample in the *psychonetrics* package in R (0.7.1; Epskamp, 2020a). Once the model was estimated, nonsignificant parameters were recursively pruned at $\alpha = .01$ and then parameters were added one at a time based on modification indices to minimize the BIC, via the step-up function. This data-driven approach is consistent with network methods (Epskamp, Borsboom, & Fried, 2018a; Fried & Cramer, 2017). Given this, model fit was expected to be good, with comparative fit index (CFI) > .95 and root mean square error of approximation (RMSEA) < .06 (Hu & Bentler, 1999).

Once each full sample network was estimated (via basic pruning and stepwise modification), three matrices from the model were extracted for invariance testing and further consideration: (a) the temporal matrix which encodes directed partial correlations for the average within-person effects over time; (b) the contemporaneous matrix which encodes partial correlations for the average within-person effects within lags (after accounting for the temporal effects); (c) the between-persons matrix which encodes partial correlations for stable trait-like differences across all time points. Average networks across all 32 models, excluding edges that occurred less than 50% of the time following Lin, Fried, and Eaton (2020), were plotted in *qgraph* (1.6.5; Epskamp et al., 2012) with red lines indicating negative parameter values (edges) and blue positive. In the temporal network, arrows between nodes indicate directed partial correlations while curved arrows represent autoregressions.

Finally, strength centrality was considered for networks in each model. Strength represents the sum of absolute edge weights for any given node (Costantini et al., 2015). For temporal networks, this includes both in-strength and out-strength, with the former indicating the relative predictability and the latter the relative influence of the target node. For undirected networks, a single strength index represents the overall extent to which a given node is directly influenced by or influences others.

Network matrices were also inspected to determine the number and size of edges and whether these were in expected directions.

Gender invariance

Following standard practices for invariance testing two models were tested: an unconstrained model or H1, and a constrained model or H0, where H0 is nested in H1. Temporal, between, and contemporaneous matrices were used to determine which parameters should be considered in an unconstrained model (i.e., those retained in the whole sample were estimated for each group). In this model these parameters of interest were freely estimated in girls and boys simultaneously to provide a point of comparison for subsequent constraints. In the constrained model, all three matrices were then set to equality in girls and boys, and the resulting model was compared to the unconstrained model. Given the sample size of the current study, models were compared based on the Akaike information criterion (AIC) and BIC which penalize for model complexity (van de Schoot, Lugtig, & Hox, 2012), rather than chi-square difference testing which can be sensitive to large samples (Crede & Harms, 2019). Lower values for AIC and BIC indicate better model fit. Since the constrained model was more parsimonious, we interpreted higher AIC and BIC values for the constrained model as indicative of noninvariance.

Results

Gender was missing for .3% of the sample. Missing data for survey indicators were low for the first wave but higher for subsequent time points (Time 1 = 2.6%–6.6%, Time 2 = 16.4%–20.9%, Time 3 = 25.9%–29.6%). Descriptive statistics are summarized in Table 1. The average fit of models is presented in Table 2 and a full summary of fit statistics for each model can be found in the supplementary material (S5). In general, differences between equivalent datasets using different estimation algorithms were small indicating good stability across these. Though data-driven approaches were used to estimate models, and parameter estimates varied, the stable good fit across conditions nevertheless indicated that stationarity constraints imposed in the model (paths a and b in Figure 1) were reasonable in all cases (Epskamp, 2020b). In terms of invariance, the same mixed result was found across all conditions: AIC favored the unconstrained model while BIC favored the constrained model. This suggests differences in network structure between girls and boys were likely small. Post-hoc consideration of RMSEA and CFI also revealed differences typically considered to be small (Meade, Johnson, & Braddy, 2008),[2] (−.007 to −.006 for CFI, with $M = -.006$, $SD < .001$; range within <−.001 for RMSEA, $M < .001$, $SD < .001$).

Edges for contemporaneous and between networks are interpreted as partial correlation coefficients, and those for temporal as directed partial directed correlations (standardized beta coefficients). For each network within each condition the number of parameters, means, standard deviations, and number of negative edges (unexpected results relative to our hypotheses, given the recoding of indicators to have a positive manifold) can be seen in the supplementary material (S6). Between networks had the fewest edges (3–23), though these were relatively large (ranging in absolute value from $r = .004$–$r > .99$ for similar indicators such as *distracted* and *restless* in some models with the mean of mean edge sizes within networks across conditions $M = .40$). No unexpected negative edges were found for between networks in any condition. Contemporaneous networks were more densely connected with 116–130 edges ($r = .06$–.27 in absolute value with the mean of mean edge sizes within networks across conditions $M = .07$), and with consistent unexpected negatives across all conditions (7–12; for example, a small negative edge featured in every contemporaneous network for the being bullied and [not] think clearly indicators). Temporal networks were also dense (166–196 edges; β = .05–.27 in absolute value, mean of means $M = .06$) with 1–3 unexpected edges found for each condition. These were consistently found for *worry* → *think*, *school* → *withdrawn*, and *unhappy* → *peer* (this was nonsignificant in eight conditions). Most estimated parameters were significant across conditions (0–12 were nonsignificant for any given condition, $p < .01$). In terms of edge parameters, only temporal networks occasionally included nonsignificant edges: 19 different edges in temporal networks were nonsignificant in different conditions, with most of these edges not occurring frequently across conditions or only rarely being nonsignificant (full information for all parameters in all conditions is provided in the supplementary material (S7–S10).

Spearman correlations between weight matrices for networks of the same type (e.g., temporal or contemporaneous) were high: Between M ρ = .82 SD ρ = .14; contemporaneous M ρ = .88 SD ρ = .06; temporal M ρ = .91 SD ρ = .03, suggesting networks were similar across conditions (full correlation matrices can be seen in the supplementary material, S11).

To summarize these networks across all conditions, for each of between, contemporaneous, and temporal networks, the mean of edges was calculated after excluding those that appeared in less than 50% of conditions. This resulted in 182 edges (37.60% of all possible edges) being retained in the average temporal network, all of which appeared across all conditions. Similar stability was seen for the between and contemporaneous networks, with all edges estimated across conditions appearing in 50% or more conditions (between: six edges, 2.60%; contemporaneous: 134 edges, 58.01%). The mean edge size for the average between network was $r = .32$ (range = .52), $r = .06$, (range = .32) for the average contemporaneous network, and $r = .06$ (range = .29) for the average temporal network. Autoregressive effects were present for all nodes in the average temporal network and ranged from .06 to .24 ($M = .14$, $SD = .04$). The average networks are summarized in Figure 2 with the thickness of edges scaled across the three panels (i.e., it is equivalent across each plot), and the supplementary material (S12). As mentioned above, a handful of nonsignificant parameters were found in temporal networks, six of which appear in the averaged network (all were nonsignificant only once across the 32 conditions, except unhappy → peer as described above).

Strength centrality was calculated for temporal and contemporaneous networks only, given the sparsity found for the between networks. Which nodes were most central, tended to depend on the condition. In and out strength for the temporal networks are shown in Figures 3 and 4, while strength for the contemporaneous network is shown in Figure 5. *Stress* was consistently high for in-strength but other nodes varied substantially. *Stress* was again fairly consistently one of the most central for out-strength as was *worry*, though again substantial variation in out-strength was seen for most nodes. *Worry* and *think* were consistently the most central for strength in the contemporaneous network. Nodes that were represented by varying items, depending on the condition, often showed particular discrepancies for strength (e.g., *mind* in Figure 3). However, nodes with the same item across conditions also showed substantial variation (e.g., *worry*

[2] We are not aware of simulation work providing recommendations for the size of alternative fit index differences for network invariance and therefore provide this example for confirmatory factor analysis (which recommends CFI difference of <.002 to consider invariance) since it includes larger samples closest to that used here.

Table 2. Average fit across datasets by model type

	χ^2		df		CFI		RMSEA		AIC		BIC	
Model type	*M* (*SD*)	Min/max	*M* (*SD*)	Min/max	*M* (*SD*)	Min/max	*M* (*SD*)	Min/max	*M* (*SD*)	Min/max	*M* (*SD*)	Min/max
Unpruned	7364.94 (314.00)	6902.29/ 8028.15	1265.00 (0.00)	1265/ 1265	0.979 (0.001)	0.976/ 0.980	0.017 (0.000)	0.017/ 0.018	2180850.22 (26288.05)	2144674.95/ 2218121.15	2188612.75 (26288.05)	2152437.48/ 2225883.68
Pruned												
(*p* = .01)	13282.74 (529.08)	12288.01/ 14383.54	1906.88(9.02)	1892/ 1919	0.960 (0.002)	0.956/ 0.964	0.019 (0.000)	0.019/ 0.020	2185484.26 (26280.05)	2148923.91/ 2223074.19	2188323.30 (26280.96)	2151854.03/ 2225866.25
Stepup	12576.36 (1046.90)	10332.04/ 13593.12	1901.44 (14.94)	1865/ 1917	0.963 (0.003)	0.958/ 0.969	0.019 (0.001)	0.017/ 0.020	2184788.76 (26008.62)	2148429.69/ 2222664.55	2187669.51 (26033.41)	2151367.49/ 2225471.95
Pruned unconstrained	15812.44 (472.83)	14947.65/ 16875.93	3813.75 (18.05)	3784/ 3838	0.957 (0.002)	0.953/ 0.960	0.020 (0.000)	0.019/ 0.021	2175247.69 (26200.58)	2139017.27/ 2212536.20	2180923.85 (26202.95)	2144875.54/ 2218118.42
Stepup unconstrained	15249.43(948.84)	13049.01/ 16226.61	3802.88 (29.88)	3730/ 3834	0.959 (0.003)	0.954/ 0.965	0.020 (0.001)	0.018/ 0.020	2174706.42 (25933.39)	2138772.35/ 2212236.48	2180465.97 (25984.64)	2144645.95/ 2217849.37
Pruned constrained	17808.36 (477.39)	16983.10/ 18881.08	4117.88 (9.02)	4103/ 4130	0.951 (0.002)	0.946/ 0.954	0.021 (0.000)	0.020/ 0.021	2176635.36 (26235.31)	2140363.03/ 2213952.87	2179979.52 (26236.43)	2143798.24/ 2217250.06
Stepup constrained	17264.63 (918.01)	15192.08/ 18246.28	4112.44 (14.94)	4076/ 4128	0.953 (0.003)	0.948/ 0.959	0.020 (0.001)	0.019/ 0.021	2176102.51 (25966.31)	2140120.64/ 2213664.02	2179488.36 (25991.75)	2143563.53/ 2216976.55

Note. df = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation; AIC = Akaike information criterion; BIC = Bayesian information criterion; *M* = mean; *SD* = standard deviation.

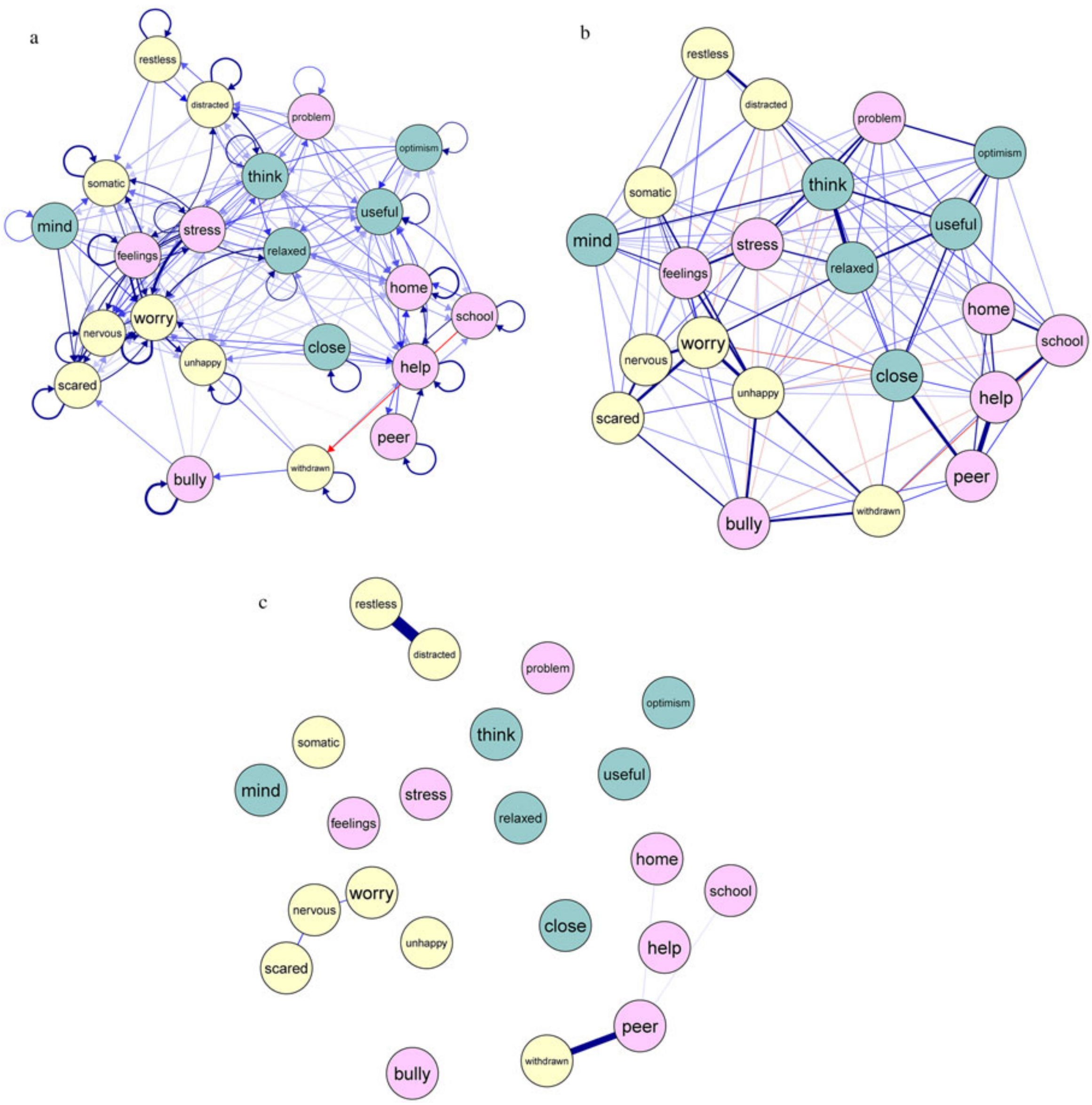


Figure 2. Average networks across conditions. Panel A: average temporal network. Panel B: average contemporaneous network. Panel C: average between network.

in Figure 3), given the conditional nature of edges which account for all others in the model.

Each line represents how the in-strength of each node varies depending on the condition. Only nodes where the maximum in-strength is always >.40 are shown in color and labeled for ease of reading.

Each line represents how the out-strength of each node varies depending on the condition. Only nodes where the maximum out-strength is always >.50 are shown in color and labeled for ease of reading.

Each line represents how the strength of each node varies depending on the condition. Only nodes where the maximum strength is always >.90 are shown in color and labeled for ease of reading.

Discussion

We explored stable trait-like and within-person associations over time for internalizing symptoms, well-being and inter and intrapersonal correlates at the indicator level. A multiverse approach was adopted, varying estimation algorithms and operationalizations of certain indicators, given that secondary data analysis was conducted, and new methods were used (Epskamp, 2019; Weston et al., 2019). Though network analyses have boomed in recent years (Robinaugh, Hoekstra, Toner, & Borsboom, 2020), this was the first study to adopt a crossed multiverse design, to our knowledge, and an early example of Epskamp's (2020b) panel methodology. While previous work has considered longitudinal relationships between internalizing symptoms and inter/intrapersonal correlates (e.g., Goodman, Samek, Wilson, Iacono, & McGue, 2019; Saint-Georges & Vaillancourt, 2020), work at the indicator level was lacking. This revealed relationships between indicators of different domains, suggesting latent-variable approaches may miss complexity. Similarly, while some work has considered both symptoms and well-being over time (e.g., Patalay & Fitzsimons, 2018), covariance between these domains was only considered by controlling for each at the first time point.

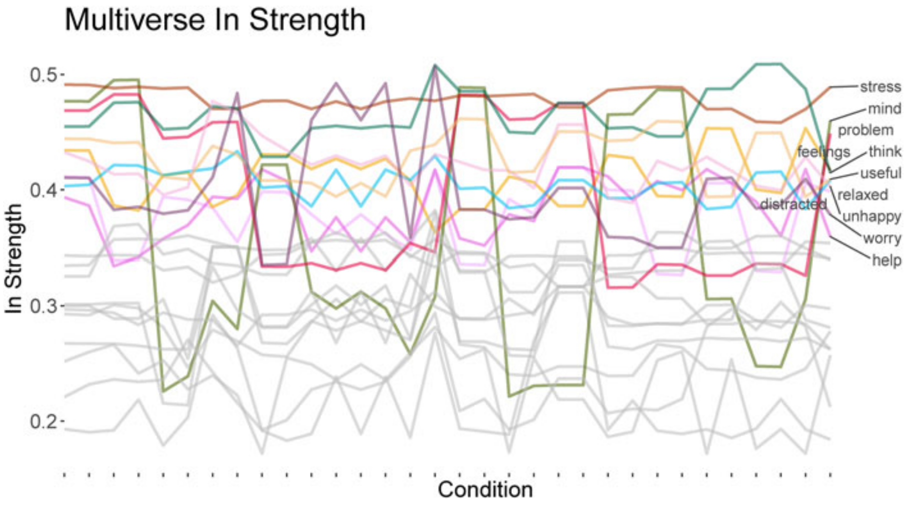


Figure 3. In-strength for each temporal network. Each line represents how the in-strength of each node varies depending on the condition. Only nodes where the maximum in-strength is always >.40 are shown in color and labeled for ease of reading.

We found a sparse between-person network with few strong associations, while the contemporaneous (average within-lag within-person associations) and temporal (directed within-person associations) networks, were densely connected. All weights matrices were highly correlated, and networks showed good stability across conditions. We did not find clear evidence that networks differed between girls and boys, and results were consistent across conditions. Findings suggest if differences existed for the indicators used here, they were likely trivial. Finally, the choice of item operationalization had a substantial impact on strength centrality (considered for the within-person networks), though certain nodes were consistently central.

Between-person Findings

The between network revealed partial correlations in expected directions, some of which were very large. These were between attentional, anxiety, and social indicators. These could reflect consistent cognitive vulnerabilities, environments, personality traits (e.g., agreeableness and neuroticism) or stable biological factors (Fraley & Roberts, 2005). There were notably no between-person relationships among indicators of different domains (e.g., internalizing and well-being or internalizing and social correlates) despite the fact that such domains have shown meaningful relationships elsewhere (e.g., Patalay & Fitzsimons, 2018). The contrast in our findings with prior work could result for several reasons, including our disaggregation of within and between-person effects, control of informant-type, and separate modeling of temporal and contemporaneous effects.

Though there was a relatively strong effect between peer support and withdrawal, which could be considered different domains (internalizing and interpersonal), we interpret this in line with the other effects in the between network: Those indicators involved were very similar and tended to be rated in similar ways over time, that is, a trait-like tendency over time to rate high or low peer support was strongly related to a trait-like tendency to rate low or high social withdrawal. The fact that indicators of different domains were conditionally independent in the between network suggests that covariance between these domains may be more state like. We were able to identify this by controlling for trait-like reporting effects in the between network. The relative sparsity of the between network also indicates the majority of covariances were not stable and trait-like, consistent with the rapidly changing developmental context of early adolescence described in the introduction.

Within-person Findings

Dense within-person, temporal and contemporaneous, networks were found. These findings fit with systems approaches in which aspects from different levels (e.g., home and intrapersonal factors) interact with one another (Bronfenbrenner, 2005; Evans et al., 2013; Masten & Barnes, 2018). Furthermore, there was little evidence of particular associations for certain inter or intrapersonal factors being associated with only symptoms or well-being as has been suggested elsewhere (Patalay & Fitzsimons, 2016). Rather, symptoms, well-being and inter/intrapersonal factors seemed to influence one another in similar ways.

While both within-person networks were relatively dense, larger relationships were typically seen in the contemporaneous network. The current study sought to understand relationships between specific indicators (e.g., thinking clearly and being bullied) rather than latent constructs (e.g., well-being or peer problems). While levels of specific indicators such as these likely have meaningful relationships over time, the dense contemporaneous network suggests that interactions between the indicators modeled here often happened more quickly than annually (Epskamp et al., 2018a,b). Since both contemporaneous and temporal networks were relatively dense, many edges were common across both of these networks. Our results therefore suggest that indicators influenced one another both within and across lags. This further points to rapid changes in mental health and correlate variables, consistent with the rapid social, physical, and psychological development seen in early adolescence (Dahl et al., 2018; Patton et al., 2016; Sawyer et al., 2018). To better understand how these processes unfold, future work should vary the length between study waves, and there is a particular need for work focusing on shorter intervals.

Some of the larger effects in the temporal network were autoregressions, with each node showing such an effect. While the indicators studied here are known to be stable or show increasing trajectories (Meeus, 2016), meaning autoregressive effects would be expected, this finding is noteworthy. First, our analysis was at the indicator level, suggesting stability or reinforcement of these states can be specific to this level, rather than the domain (e.g., internalizing symptoms). Second, while latent-variable

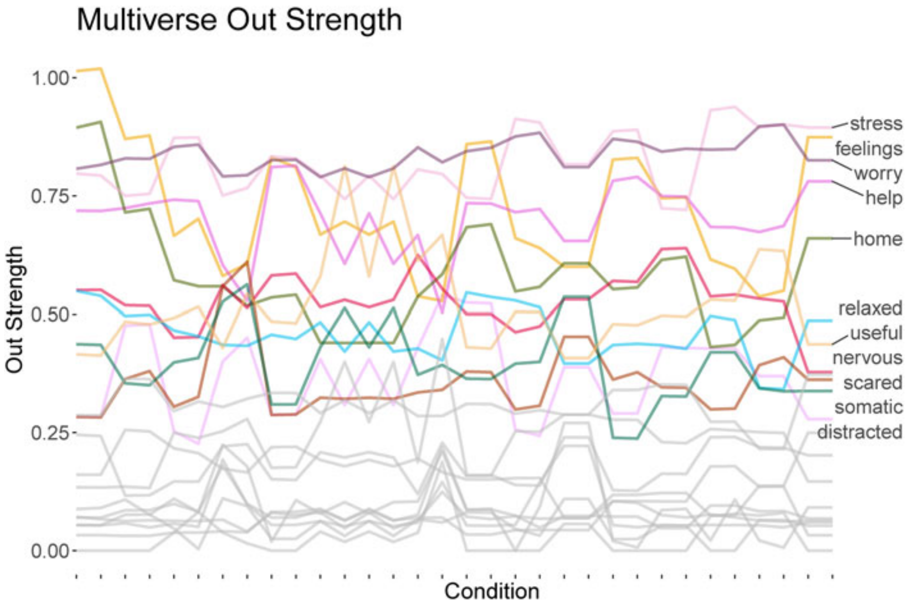


Figure 4. Out-strength for each temporal network. Each line represents how the out-strength of each node varies depending on the condition. Only nodes where the maximum out-strength is always >.50 are shown in color and labeled for ease of reading.

approaches account for construct-level covariance, parameters in our model controlled for those to *all* other indicators, and thus also included unique variance beyond that explained by a potential latent variable, which could be substantively important. Third, autoregressive parameters in our analysis accounted for stable between-person differences over time, thus representing more accurate within-person reinforcement of individual experiences over time. These within-person autoregressions have been interpreted by some as warning signals for transition into more disordered states (e.g., van de Leemput et al., 2014), but since we examined a large cohort via survey methods, we did not consider whether individuals were more or less disordered over time.[3] Nevertheless, the age range studied here is thought to be critical in the emergence of mental health problems (Jones, 2013) and rates are known to increase in this age range (Merikangas et al., 2010; NHS Digital, 2018). It may be therefore that cementing of symptoms, well-being indicators and inter/intrapersonal factors all contribute to this change.

Edges were mostly in expected directions, relative to our hypotheses. However, unexpected negative parameters were observed consistently in the temporal and contemporaneous networks. A certain level of such effects in partial correlation networks could be consistent with the nominal alpha level, or due to conditioning on common effects (Epskamp, Waldorp, Mõttus, & Borsboom, 2018b). We are therefore cautious in providing substantive interpretation of these results. However, one such effect was particularly stable across conditions and temporal and contemporaneous networks, that between *withdrawn* and (lack of) school support.[4] While we anticipated that internalizing symptoms would be positively associated with perceived lack of social support, it may be that adolescents who reported feeling socially withdrawn were focusing on the peer level when responding to the *withdrawn* item. In fact, the full item reads “I am usually on my own. I generally play alone or keep to myself”. It is possible that adolescents who felt withdrawn from their peers tended to garner more support from, or were dependent on, school staff as can be the case for loneliness (Galanaki & Vassilopoulou, 2007).

Centrality

Strength centrality appeared more stable for the contemporaneous network than the temporal. *Think* and *unhappy* had the highest strength, depending on the condition, followed by *worry*, while the rank order of strength varied for the remaining nodes. This suggests that when considering relationships that happened more quickly than over a year, feeling unhappy and thinking clearly were particularly connected to other indicators, sharing the most variance with others (Costantini et al., 2015). Internalizing symptom and well-being indicators appeared to be among the most important in the contemporaneous network, suggesting both outcomes are intricately connected to each other and correlates. This further supports the use of well-being measures to better understand internalizing states, since well-being indicators clearly shared meaningful variance with other indicators without being redundant with respect to internalizing indicators.

Being able to deal with stress was one of the most consistently strongly predicted and influential nodes, suggesting that for effects that happened over the course of a year, the indicators in the model often related to this outcome via relatively strong directed partial correlations. Conversely, finding it hard to control feelings, an item designed to measure the same underlying trait as the *stress* indicator, was sometimes the most central for out-strength, while at other times several other indicators were stronger, and substantially lower values were seen. *Worry*, which was fairly consistently one of the most central nodes across conditions for out-strength, varied substantially for in-strength. Other particularly wide variations for in-strength were seen for the *mind* and *problem* indicators, both of which had varying operationalizations across conditions.

Given the finding that the contemporaneous network remained dense, we do not interpret only the temporal centrality results as indicative of risk factors or outcomes. Rather, results

[3] This would have relied on total scores which can be problematic (McNeish & Wolf, 2020) and inconsistent with our modeling approach.

[4] The node is considered as lack of school support due to the recoding prior to analysis to obtain a positive manifold for the easy detection of results counter to hypotheses across conditions.

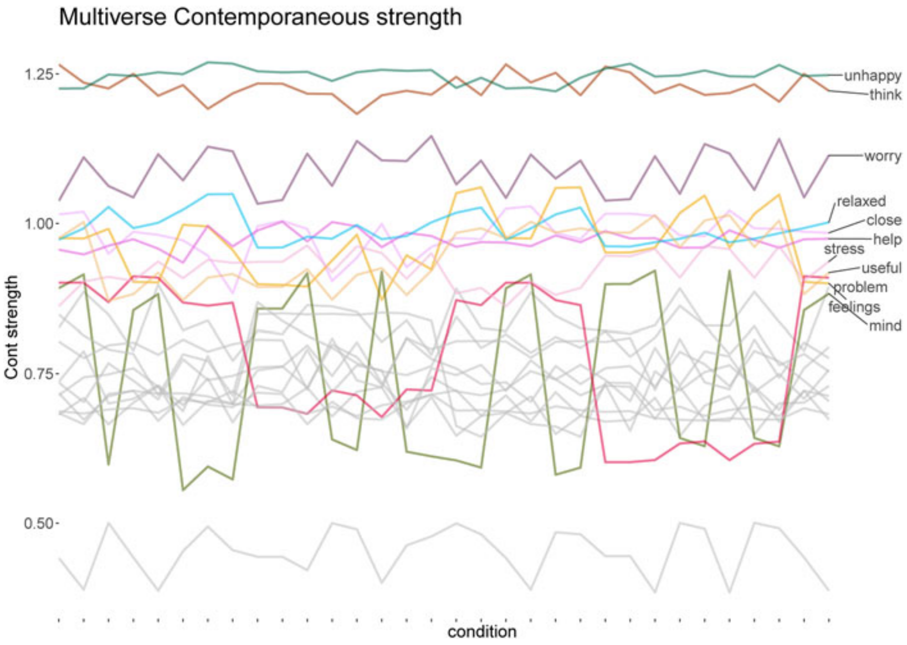


Figure 5. Contemporaneous strength. Each line represents how the strength of each node varies depending on the condition. Only nodes where the maximum strength is always >.90 are shown in color and labeled for ease of reading.

suggest that worry, managing stress, thinking clearly and unhappiness may be key indicators for the development of adolescents' mental health. While more work is needed, this suggests that worry and unhappiness may be particularly important symptoms in early adolescence when considering how rapid developmental change is navigated. In turn, the *think* and *stress* indicators' centrality suggest that such cognitive indicators may play an important role in the reinforcement of social and psychological processes in this age group.

Our findings also highlight the importance of which items are chosen, and the issues of measurement error in adolescent mental health data. The stability of networks across samples using the same items has been given attention in recent years (e.g., Borsboom et al., 2017; Forbes, Wright, Markon, & Krueger, 2017), as has the stability across different measures in certain fields (Fried et al., 2018). However, this was the first study, to our knowledge, to consider the sensitivity of network parameters to item operationalizations in the *same* sample. We found that while some nodes showed relative stability others varied in strength centrality for both indicators that were constant and those that varied across conditions.

Gender invariance

Gender invariance results were stable across conditions but, we were unable to determine clear support for invariance based on AIC and BIC as recommended by van de Schoot et al. (2012). Consistent with known possible behavior of these criteria, AIC favored the model with more parameters (unconstrained), while BIC did the opposite, favoring the constrained model (Vrieze, 2012). Since we had no clear rationale to favor one over the other, we consider these results in light of other literature and indices (post-hoc). Kan, van der Maas, and Levine (2019) found the same pattern of AIC and BIC for their unconstrained and constrained networks. They concluded that the same structure was applicable to both groups, though at least one edge varied in magnitude. Our post-hoc consideration of CFI and RMSEA also suggested trivial differences, thus supporting the approximate invariance of networks between boys and girls. Substantively, a single pattern of edges fitted both girls and boys, though small differences may exist in the strength of different relationships between particular nodes. Results should be replicated considering other measures and samples, but this suggests tentative evidence that girls and boys may experience quantitative rather than qualitative differences in risk and protective factors for internalizing symptoms, when considering inter/intrapersonal correlates.

Implications

Taken together, the strong dissociated relationships at the between level and densely connected nodes at the within level suggest the apparent discriminant validity of scales may particularly capture between-person differences rather than profiles within individuals. This is consistent with the fact that measures are typically developed using between-person (i.e., cross-sectional) data, such that the covariance structure from which the model is estimated describes variation between people (Molenaar, 2004). Though many analyses assume a blend of within and between effects is modeled without explicitly attending to this, it is often the case that within and between associations are not aligned (Curran & Bauer, 2011). Future work should therefore consider further the within and between properties of measures such as those used here, since they are typically used to probe within-person effects.

Findings further suggest integrated indicator-level approaches to adolescents' mental states and perceived resources should be considered, rather than testing to diagnose specific disorders. Our analysis therefore represents an example of how clinical and research approaches can better align, as has been pointed out for network approaches more generally (Borsboom, 2017). While formulations are often preferred over strict diagnostic criteria by clinicians (Johnstone, 2018), research has tended to rely on simplistic total scores or latent variables to define groups and categories. These are powerful approaches, with many advantages such as the estimation of measurement error. Nevertheless, as indicator-level approaches gain increasing attention (Robinaugh et al., 2020), analyses such as ours can offer more detailed insights. While much more work is needed, the current study demonstrates that brief surveys deployed in large samples can be modeled in more nuanced ways. There is therefore potential to move beyond disorder-level (i.e., total-score) approaches. In addition, it may not be enough to disaggregate within and between effects at the construct level, since within-person effects likely happen across domains in a complex way (Borsboom et al., 2018). More transdiagnostic and indicator-level work is therefore needed to better understand within and between-person effects.

In addition to the substantive implications, our multiverse design revealed methodological issues. Where observed-data level networks are considered, as is typically the case (Robinaugh et al., 2020), rather than at the latent level (Epskamp, 2020b; Epskamp et al., 2017), researchers should be aware that item-level error may affect conclusions. Since many authors rely on single measures of each construct in their datasets, they will be unable to verify whether, for instance, centrality is robust to variations in items. Our out-strength results particularly demonstrate that had we chosen any one of the 32 conditions as the focus of our analysis, our conclusions could have varied substantially. While there are calls for increased use of latent networks, our approach also reveals that even in large rich datasets, there may not be enough indicators of each construct to conduct such analysis. For instance, our dataset had only one bullying item. We therefore echo the recent call for methods to be designed explicitly with network methods in mind (Robinaugh et al., 2020).

Strengths and limitations

The current study drew on a large sample, disaggregated within-person variance from stable trait-like effects, and incorporated a comprehensive multiverse design. Despite this a number of limitations must be acknowledged. First, the panel methodology adopted did not allow us to control for stable covariates, such as socioeconomic status. While the sample was purposively drawn to target those at risk, and therefore generally consisted of more deprived adolescents, there was variation in this. The sample was therefore also not representative and results should only be generalized to similar community samples with above average levels of deprivation.

We were also unable to account for the nonnormal ordinal nature of our data since this is not yet possible in *psychonetrics*. Nevertheless, this is consistent with much of the network literature to date, which often treats similar Likert-type data to that used here as continuous (Robinaugh et al., 2020). In addition, the use of a polychoric matrix to account for the ordinal nature of items in skewed data, such as that used here, can lead to bias (Fried, van Borkulo, & Epskamp, 2020). It can also lead to convergence issues in samples with substantial missingness, as was the case here, suggesting FIML was more appropriate. We also had little to draw on to interpret our invariance analyses, and more work is needed to understand the properties of fit indices when comparing networks.

Quality issues have been highlighted for some SDQ items, from which internalizing and bullying indicators were drawn (Black et al., 2020a), though self-report mental health measures are typically of low quality (Bentley et al., 2019). Finally, decisions about which items were interchangeable were subjectively considered based on content, though decisions in multiverse analyses are not expected to be uniform across researchers (Simonsohn et al., 2020).

Conclusion

The current multiverse panel network model allowed consideration of complex interactions between indicators of mental health and inter/intrapersonal factors consistent with theory and clinical approaches (Bronfenbrenner, 2005; Johnstone, 2018). Stable trait-like effects within anxiety, attentional and social indicators were found that were insensitive to analytical decisions. No clear differences were observed between boys and girls. Within-person networks were densely connected and relationships between indicators often unfolded within waves, suggesting more work should consider shorter lags. Mental health and inter/intrapersonal indicators appeared to relate to one another in similar complex ways. Our multiverse design revealed that the particular operationalization of items can have substantial effects on conclusions. Nevertheless, indicators such as thinking clearly, unhappiness, dealing with stress and worry showed more consistent centrality, suggesting these indicators may play particularly important roles in the development of mental health in adolescence.

Supplementary Material. The supplementary material for this article can be found at https://doi.org/10.1017/S0954579421000225

Data Availability Statement. The HeadStart (HS) survey data on mental health and well-being belongs to the Evidence Based Practice Unit (a collaboration between UCL and the Anna Freud National Centre for Children and Families, AFNCCF), who led the HS evaluation. The authors accessed this survey data via membership in a consortium involved with the HS evaluation. As collaborators on the main HS evaluation, the authors were granted secure remote access to this data by the principal investigator of the main HS evaluation, Dr. Jessica Deighton. HS data cannot be made publicly available, since consent was not obtained from participants for the public sharing of their survey responses. However, an anonymized version of the survey dataset used in the present paper is available on request from Dr. Jessica Deighton (Jessica.DeightonPhD@annafreud.org) or Dr. Tanya Lereya (Tanya.lereya@annafreud.org) under the following terms: (a) schedule and arrange for site visit to AFNCCF to analyze data (password to user account supplied); (b) analysis to be worked on in situ; (c) results (but not data) taken away. In the event that either of these individual leaves the AFNCCF, updated contact information for new guardians of the data will be provided to the *Journal of Abnormal Child Psychology*. Code is provided in the supplementary material.

Funding Statement. The data used in this study were collected as part of the HeadStart learning program and supported by funding from The National Lottery Community Fund, grant R118420. The content is solely the responsibility of the authors and it does not reflect the views of The National Lottery Community Fund.

Conflicts of Interest. None.

Ethical Standards. Ethical approval was provided by the University College London Ethics Committee (UCL Ref: 8097/003).

References

Alexander, L. M., Salum, G. A., Swanson, J. M., & Milham, M. P. (2020). Measuring strengths and weaknesses in dimensional psychiatry. *Journal of Child Psychology and Psychiatry*, *61*, 40–50. doi:10.1111/jcpp.13104

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th edn). Arlington, VA: American Psychiatric Association.

Antaramian, S. P., Huebner, S. E., Hills, K. J., & Valois, R. F. (2010). A dual-factor model of mental health: Toward a more comprehensive understanding of youth functioning. *American Journal of Orthopsychiatry*, *80*, 462–472. doi:10.1111/j.1939-0025.2010.01049.x

Bartels, M., Cacioppo, J. T., van Beijsterveldt, T. C. E. M., & Boomsma, D. I. (2013). Exploring the association between well-being and psychopathology in adolescents. *Behavior Genetics*, *43*, 177–190. doi:10.1007/s10519-013-9589-7

Bentley, N., Hartley, S., & Bucci, S. (2019). Systematic review of self-report measures of general mental health and wellbeing in adolescent mental health. *Clinical Child and Family Psychology Review*, *22*, 225–252. doi:10.1007/s10567-018-00273-x

Bertha, E. A., & Balázs, J. (2013). Subthreshold depression in adolescence: A systematic review. *European Child & Adolescent Psychiatry*, *22*, 589–603. doi:10.1007/s00787-013-0411-0

Black, L., Mansfield, R., & Panayiotou, M. (2020a). Age appropriateness of the self-report strengths and difficulties questionnaire. *Assessment*, *0*, 1073191120903382. doi:10.1177/1073191120903382

Black, L., Panayiotou, M., & Humphrey, N. (2019). The dimensionality and latent structure of mental health difficulties and wellbeing in early adolescence. *PLoS One*, *14*, e0213018. doi:10.1371/journal.pone.0213018

Black, L., Panayiotou, M., & Humphrey, N. (2020b). The special relationship of internalizing symptoms and wellbeing: A cross-validation study considering indicator-level associations beyond the dual-factor model of mental health. doi:10.31234/osf.io/stajk

Bor, W., Dean, A. J., Najman, J., & Hayatbakhsh, R. (2014). Are child and adolescent mental health problems increasing in the 21st century? A systematic review. *Australian & New Zealand Journal of Psychiatry*, *48*, 606–616. doi:10.1177/0004867414533834

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*, 5–13. doi:10.1002/wps.20375

Borsboom, D., Cramer, A., & Kalis, A. (2018). Brain disorders? Not really… Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, 1–54. doi:10.1017/S0140525X17002266

Borsboom, D., Cramer, A. O. J., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The small world of psychopathology. *PLoS One*, *6*, e27407. doi:10.1371/journal.pone.0027407

Borsboom, D., Fried, E. I., Epskamp, S., Waldorp, L. J., van Borkulo, C. D., van der Maas, H. L. J., & Cramer, A. O. J. (2017). False alarm? A comprehensive reanalysis of "evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger (2017). *Journal of Abnormal Psychology*, *126*, 989–999. doi:10.1037/abn0000306

Bronfenbrenner, U. (2005). *Making human beings human: Bioecological perspectives on human development*. Thousand Oaks, USA: Sage.

Carrellas, N. W., Biederman, J., & Uchida, M. (2017). How prevalent and morbid are subthreshold manifestations of major depression in adolescents? A literature review. *Journal of Affective Disorders*, *210*, 166–173. doi:10.1016/j.jad.2016.12.037

Collishaw, S. (2015). Annual research review: Secular trends in child and adolescent mental health. *Journal of Child Psychology and Psychiatry*, *56*, 370–393. doi:10.1111/jcpp.12372

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the art personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13–29. doi:10.1016/j.jrp.2014.07.003

Costantini, G., Richetin, J., Preti, E., Casini, E., Epskamp, S., & Perugini, M. (2019). Stability and variability of personality networks. A tutorial on recent developments in network psychometrics. *Personality and Individual Differences*, *136*, 68–78. doi:10.1016/j.paid.2017.06.011

Crede, M., & Harms, P. (2019). Questionable research practices when using confirmatory factor analysis. *Journal of Managerial Psychology*, *34*, 18–30. doi:10.1108/JMP-06-2018-0272

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, *62*, 583–619. doi:10.1146/annurev.psych.093008.100356

Dahl, R. E., Allen, N. B., Wilbrecht, L., & Suleiman, A. B. (2018). Importance of investing in adolescence from a developmental science perspective. *Nature*, *554*, 441. doi:10.1038/nature25770

Dedrick, R. F., Greenbaum, P. E., Friedman, R. M., Wetherington, C. M., & Knoff, H. M. (1997). Testing the structure of the child behavior checklist/4-18 using confirmatory factor analysis. *Educational and Psychological Measurement*, *57*, 306–313. doi:10.1177/0013164497057002009

Deighton, J., Lereya, S., Casey, P., Patalay, P., Humphrey, N., & Wolpert, M. (2019). Prevalence of mental health problems in schools: Poverty and other risk factors among 28 000 adolescents in England. *British Journal of Psychiatry*, *215*(3), 565–567. doi:10.1192/bjp.2019.19

Demkowicz, O., Panayiotou, M., Ashworth, E., Humphrey, N., & Deighton, J. (2019). The factor structure of the 4-item perceived stress scale in English adolescents. *European Journal of Psychological Assessment*, *36*, 913–917. doi:10.1027/1015-5759/a000562

Department for Education. (2017a). Pupil premium: allocations and conditions of grant 2016 to 2017. Retrieved from https://www.gov.uk/government/publications/pupil-premium-conditions-of-grant-2016-to-2017

Department for Education. (2017b). Schools, pupils and their characteristics: January 2017. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650547/SFR28_2017_Main_Text.pdf

Department for Education. (2017c). Special educational needs in England: January 2017. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633031/SFR37_2017_Main_Text.pdf

Epskamp, S. (2019). Reproducibility and replicability in a fast-paced methodological world. *Advances in Methods and Practices in Psychological Science*, *2*, 145–155. doi:10.1177/2515245919847421

Epskamp, S. (2020a). *Package 'psychonetrics'*. Retrieved from https://cran.r-project.org/web/packages/psychonetrics/psychonetrics.pdf

Epskamp, S. (2020b). Psychometric network models from time-series and panel data. *Psychometrika*, *85*, 206–231. doi: 10.1007/s11336-020-09697-3

Epskamp, S., Borsboom, D., & Fried, E. I. (2018a). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*, 195–212. doi:10.3758/s13428-017-0862-1

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). Qgraph: Network visualizations of relationships in psychometric data. 2012. *Journal of Statistical Software*, *48*, 18. doi:10.18637/jss.v048.i04

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent Variable models. *Psychometrika*, *82*, 904–927. doi:10.1007/s11336-017-9557-x

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018b). The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*, 453–480. doi:10.1080/00273171.2018.1454823

Evans, G. W., Li, D., & Whipple, S. S. (2013). Cumulative risk and child development. *Psychological Bulletin*, *139*, 1342. doi:10.1037/a0031808

Flouri, E., Papachristou, E., Midouhas, E., Ploubidis, G. B., Lewis, G., & Joshi, H. (2019). Developmental cascades of internalising symptoms, externalising problems and cognitive ability from early childhood to middle adolescence. *European Psychiatry*, *57*, 61–69. doi:10.1016/j.eurpsy.2018.12.005

Forbes, M. K., Wright, A. G. C., Markon, K. E., & Krueger, R. F. (2017). Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology*, *126*, 969–988. doi:10.1037/abn0000276

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review*, *112*, 60–74. doi:10.1037/0033-295X.112.1.60

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*, 271–288. doi:10.1080/1047840X.2020.1853461

Fried, E. I., & Cramer, A. O. J. (2017). Moving forward: Challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science*, *12*, 999–1020. doi:10.1177/1745691617705892

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Huisman-van Dijk, H. M., Bockting, C. L. H., … Karstoft, K.-I. (2018). Replicability and generalizability of posttraumatic stress disorder (PTSD) networks: A cross-cultural multisite study of PTSD symptoms in four trauma patient samples. *Clinical Psychological Science*, *6*, 335–351. doi:10.1177/2167702617745092

Fried, E. I., Nesse, R. M., Zivin, K., Guille, C., & Sen, S. (2014). Depression is more than the sum score of its parts: Individual DSM symptoms have different risk factors. *Psychological Medicine*, *44*, 2067–2076. doi:10.1017/S0033291713002900

Fried, E. I., van Borkulo, C. D., & Epskamp, S. (2020). On the importance of estimating parameter uncertainty in network psychometrics: A response to Forbes et al. (2019). *Multivariate Behavioral Research*, 1–6. doi:10.1080/00273171.2020.1746903

Fritz, J., de Graaff, A. M., Caisley, H., van Harmelen, A.-L., & Wilkinson, P. O. (2018). A systematic review of amenable resilience factors that moderate and/or mediate the relationship between childhood adversity and mental health in young people. *Frontiers in Psychiatry*, *9*, 230. doi:10.3389/fpsyt.2018.00230

Fritz, J., Stochl, J., Fried, E. I., Goodyer, I. M., van Borkulo, C. D., Wilkinson, P. O., & van Harmelen, A. L. (2019). Unravelling the complex nature of resilience factors and their changes between early and later adolescence. *BMC Medicine*, *17*, 203. doi:10.1186/s12916-019-1430-6

Galanaki, E. P., & Vassilopoulou, H. D. (2007). Teachers and children's loneliness: A review of the literature and educational implications. *European Journal of Psychology of Education*, *22*, 455. doi:10.1007/BF03173466

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*, 1337–1345. doi:10.1097/00004583-200111000-00015

Goodman, R., Meltzer, H., & Bailey, V. (1998). The strengths and difficulties questionnaire: A pilot study on the validity of the self-report version. *European Child & Adolescent Psychiatry*, *7*, 125–130. doi:10.1007/s007870050057

Goodman, R. J., Samek, D. R., Wilson, S., Iacono, W. G., & McGue, M. (2019). Close relationships and depression: A developmental cascade approach. *Development and Psychopathology*, *31*, 1451–1465. doi:10.1017/S0954579418001037

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*, 102–116. doi:10.1037/a0038889

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. doi:10.1080/10705519909540118

Hyde, J. S., Mezulis, A. H., & Abramson, L. Y. (2008). The ABCs of depression: Integrating affective, biological, and cognitive models to explain the emergence of the gender difference in depression. *Psychological Review*, *115*, 291–313. doi:10.1037/0033-295X.115.2.291

Johnstone, L. (2018). Psychological formulation as an alternative to psychiatric diagnosis. *Journal of Humanistic Psychology*, *58*, 30–46. doi:10.1177/0022167817722230

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

Kalisch, R., Cramer, A. O. J., Binder, H., Fritz, J., Leertouwer, I., Lunansky, G., … van Harmelen, A.-L. (2019). Deconstructing and reconstructing resilience: A dynamic network approach. *Perspectives on Psychological Science*, *14*, 765–777. doi:10.1177/1745691619855637

Kan, K.-J., van der Maas, H. L. J., & Levine, S. Z. (2019). Extending psychometric network analysis: Empirical evidence against g in favor of mutualism? *Intelligence*, *73*, 52–62. doi:10.1016/j.intell.2018.12.004

Kaufman, T. M. L., Kretschmer, T., Huitsing, G., & Veenstra, R. (2020). Caught in a vicious cycle? Explaining bidirectional spillover between parent-child relationships and peer victimization. *Development and Psychopathology*, *32*, 11–20. doi:10.1017/S0954579418001360

Lereya, S. T., Humphrey, N., Patalay, P., Wolpert, M., Böhnke, J. R., Macdougall, A., & Deighton, J. (2016). The student resilience survey: Psychometric validation and associations with mental health. *Child and Adolescent Psychiatry and Mental Health*, *10*, 44. doi:10.1186/s13034-016-0132-5

Lin, S.-Y., Fried, E. I., & Eaton, N. R. (2020). The association of life stress with substance use symptoms: A network analysis and replication. *Journal of Abnormal Psychology*, *129*, 204–214. doi:10.1037/abn0000485

Masselink, M., Van Roekel, E., Hankin, B. L., Keijsers, L., Lodder, G. M. A., Vanhalst, J., … Oldehinkel, A. J. (2018). The longitudinal association between self-esteem and depressive symptoms in adolescents: Separating between-person effects from within-person effects. *European Journal of Personality*, *32*, 653–671. doi:10.1002/per.2179

Masten, A. S., & Barnes, A. J. (2018). Resilience in children: Developmental perspectives. *Children (Basel, Switzerland)*, *5*, 98. doi:10.3390/children5070098

Masten, A. S., & Cicchetti, D. (2010). Developmental cascades. *Development and Psychopathology*, *22*, 491–495. doi:10.1017/S0954579410000222

McElroy, E., Fearon, P., Belsky, J., Fonagy, P., & Patalay, P. (2018). Networks of depression and anxiety symptoms across development. *Journal of the American Academy of Child & Adolescent Psychiatry*, *57*, 964–973. doi:10.1016/j.jaac.2018.05.027

McElroy, E., & Patalay, P. (2019). In search of disorders: Internalizing symptom networks in a large clinical sample. *Journal of Child Psychology and Psychiatry*, *60*, 897–906. doi:10.1111/jcpp.13044

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, doi:10.3758/s13428-020-01398-0

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568–592. doi:10.1037/0021-9010.93.3.568

Meeus, W. (2016). Adolescent psychosocial development: A review of longitudinal models and research. *Developmental Psychology*, *52*, 1969–1993. doi:10.1037/dev0000243

Merikangas, K. R., He, J.-p., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., … Swendsen, J. (2010). Lifetime prevalence of mental disorders in U.S. Adolescents: Results from the national comorbidity survey replication–adolescent supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*, 980–989. doi:10.1016/j.jaac.2010.05.017

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology. *This Time Forever. Measurement: Interdisciplinary Research and Perspectives*, *2*, 201–218. doi:10.1207/s15366359mea0204_1

NHS Digital. (2018). *Mental Health of Children and Young People in England, 2017 Summary of key findings*. Retrieved from https://files.digital.nhs.uk/F6/A5706C/MHCYP%202017%20Summary.pdf

Oh, Y., Greenberg, M. T., Willoughby, M. T., Vernon-Feagans, L., Greenberg, M. T., Blair, C. B., … The Family Life Project Key, I. (2020). Examining longitudinal associations between externalizing and internalizing behavior problems at within- and between-child levels. *Journal of Abnormal Child Psychology*, *48*, 467–480. doi:10.1007/s10802-019-00614-6

Patalay, P., & Fitzsimons, E. (2016). Correlates of mental illness and wellbeing in children: Are they the same? Results from the UK millennium cohort study. *Journal of American Academic Child Adolescent Psychiatry*, *55*, 771–783. doi:10.1016/j.jaac.2016.05.019

Patalay, P., & Fitzsimons, E. (2017). *Mental ill-health among children of the new century: Trends across childhood with a focus on age 14. September 2017*. London: Centre for Longitudinal Studies.

Patalay, P., & Fitzsimons, E. (2018). Development and predictors of mental ill-health and wellbeing from childhood to adolescence. *Social Psychiatry and Psychiatric Epidemiology*, *53*, 1311–1323. doi:10.1007/s00127-018-1604-0

Patalay, P., Fonagy, P., Deighton, J., Belsky, J., Vostanis, P., & Wolpert, M. (2015). A general psychopathology factor in early adolescence. *British Journal of Psychiatry*, *207*, 15–22. doi:10.1192/bjp.bp.114.149591

Patton, G. C., Sawyer, S. M., Santelli, J. S., Ross, D. A., Afifi, R., Allen, N. B., … Viner, R. M. (2016). Our future: A lancet commission on adolescent health and wellbeing. *The Lancet*, *387*, 2423–2478. doi:10.1016/S0140-6736(16)00579-1

Petrides, K. V., & Furnham, A. (2009). *Technical manual for the trait emotional intelligence questionnaires (TEIQue)*. London: London Psychometric Laboratory.

Rapee, R. M., Oar, E. L., Johnco, C. J., Forbes, M. K., Fardouly, J., Magson, N. R., & Richardson, C. E. (2019). Adolescent development and risk for the onset of social-emotional disorders: A review and conceptual model. *Behaviour Research and Therapy*, *123*, 103501. doi:10.1016/j.brat.2019.103501

Ravens-Sieberer, U., Gosch, A., Rajmil, L., Erhart, M., Bruil, J., Duer, W., … European Kidscreen Group. (2005). KIDSCREEN-52 quality-of-life measure for children and adolescents. *Expert Review of Pharmacoeconomics & Outcomes Research*, *5*, 353–364. doi:10.1586/14737167.5.3.353

Rhemtulla, M., Cramer, A., van Bork, R., & Williams, D. R. (2018). Cross-lagged network models. 1-32.

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R., & Borsboom, D. (2020). The network approach to psychopathology: A review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*, *50*, 353–366. doi:10.1017/S0033291719003404

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*, 27–42. doi:10.1177/2515245917745629

Saint-Georges, Z., & Vaillancourt, T. (2020). The temporal sequence of depressive symptoms, peer victimization, and self-esteem across adolescence: Evidence for an integrated self-perception driven model. *Development and Psychopathology*, *32*, 975–984. doi:10.1017/S0954579419000865

Sawyer, S. M., Azzopardi, P. S., Wickremarathne, D., & Patton, G. C. (2018). The age of adolescence. *The Lancet Child & Adolescent Health*, *2*, 223–228. doi:10.1016/S2352-4642(18)30022-1

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*, 1208–1214. doi:10.1038/s41562-020-0912-z

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712. doi:10.1177/1745691616658637

Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J., & Weich, S. (2009). Internal construct validity of the Warwick-Edinburgh mental well-being scale (WEMWBS): A Rasch analysis using data from the Scottish health education population survey. *Health and Quality of Life Outcomes*, *7*, 15. doi:10.1186/1477-7525-7-15

Suldo, S., Thalji, A., & Ferron, J. (2011). Longitudinal academic outcomes predicted by early adolescents' subjective well-being, psychopathology, and mental health status yielded from a dual factor model. *The Journal of Positive Psychology*, *6*, 17–30. doi:10.1080/17439760.2010.536774

Thompson, R. A. (2019). Emotion dysregulation: A theme in search of definition. *Development and Psychopathology*, *31*, 805–815. doi:10.1017/S0954579419000282

van de Leemput, I. A., Wichers, M., Cramer, A. O. J., Borsboom, D., Tuerlinckx, F., Kuppens, P., … Scheffer, M. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, *111*, 87. doi:10.1073/pnas.1312114110

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486–492. doi:10.1080/17405629.2012.686740

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*, 228–243. doi:10.1037/a0027127

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*, *2*, 214–227. doi:10.1177/2515245919848684

WHO. (1946). *constitution of the world health organization*. Retrieved from Geneva: http://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf?ua=1

WHO. (2018). International classification of diseases for mortality and morbidity statistics (11th Revision). https://icd.who.int/browse11/l-m/en

Wolff, J. C., & Ollendick, T. H. (2006). The comorbidity of conduct problems and depression in childhood and adolescence. *Clinical Child and Family Psychology Review*, *9*, 201–220. doi:10.1007/s10567-006-0011-3

Wolpert, M., & Rutter, H. (2018). Using flawed, uncertain, proximate and sparse (FUPS) data in the context of complexity: Learning from the case of child mental health. *BMC Medicine*, *16*, 82. doi:10.1186/s12916-018-1079-6

Implications of Paper 2 for Policy Makers and/or Educators and Teachers

Building on the potential challenges/support to previous consensus on policy implications for assessment, intervention, and public systems (Iasiello & Agteren, 2020) from Paper 1, the findings of Paper 2 raise further challenges. The strong impact of operationalization (for items judged to represent the *same* indicator) on results, suggests that measurement issues should be considered more prominently. That is to say if differential effects are found for given symptom/wellbeing outcomes, measurement issues known to be endemic to adolescent mental health (Wolpert & Rutter, 2018) should be considered as an explanation. This in turn should inform the interpretation of the quality of the evidence base. This is perhaps particularly the case as the findings of similar complexity to correlates and network centrality of wellbeing and internalizing symptom indicators, suggest, in line with Paper 1, that internalizing symptoms and wellbeing are likely to behave in similar ways. This suggests the tendency to jump to the conclusion that differential intervention effects for symptoms and wellbeing are support for only one 'aspect' of mental health being affected (Iasiello & Agteren, 2020) could be flawed.

Further implications for assessment, and therefore public systems and selecting interventions, were also seen through a few consistently central indicators (thinking clearly, unhappiness, dealing with stress, and worry), which appeared more robust to measurement operationalization issues. The findings of this paper suggest these indicators should be a priority in self-report measurement (e.g., for screening or assessment and monitoring purposes).

References

Iasiello, M., & Agteren, J. v. (2020). *Mental health and/or mental illness: A scoping review of the evidence and implications of the dual-continua model of mental health*. Exeley. https://doi.org/10.3316/informit.261420605378998

Wolpert, M., & Rutter, H. (2018). Using flawed, uncertain, proximate and sparse (FUPS) data in the context of complexity: learning from the case of child mental health. *BMC Medicine*, *16*(1), 82. https://doi.org/10.1186/s12916-018-1079-6

Paper 3: Age Appropriateness of the Self-Report Strengths and Difficulties Questionnaire

Status: published, green open access (author- accepted manuscript presented in thesis)

Black, L., Mansfield, R., & Panayiotou, M. (2020). Age Appropriateness of the Self-Report Strengths and Difficulties Questionnaire. Assessment, 0(0), 1073191120903382. doi:10.1177/1073191120903382

Supplementary material can be found in Appendix 3.

Age appropriateness of the self-report Strengths and Difficulties Questionnaire

Louise Black, Rosie Mansfield and Margarita Panayiotou

The University of Manchester

Black, L., Mansfield, R., & Panayiotou, M. (2020). Age Appropriateness of the Self-Report Strengths and Difficulties Questionnaire. Assessment. https://doi.org/10.1177/1073191120903382

Author Note

Louise Black, Manchester Institute of Education, University of Manchester; Rosie Mansfield, Manchester Institute of Education, University of Manchester, Margarita Panayiotou, Manchester Institute of Education, University of Manchester.

Acknowledgments: The data used in this study were collected as part of the HeadStart learning program. The authors are therefore grateful for the work of the wider research teams at the Anna Freud Centre and the University of Manchester for their role in coordinating the evaluation, as well as collecting and managing the data. The authors also acknowledge the National Pupil Database from which demographic data were obtained. Finally, we are extremely grateful to all students who took part in this study, as well as the local authorities and schools for their help in recruiting them.

Funding: The data used in this study was collected as part of HeadStart learning program and supported by funding from the National Lottery Community Fund, grant R118420. The funders did not seek to influence any aspect of the secondary analysis reported in this study. The content is solely the responsibility of the authors and it does not reflect the views of the National Lottery Community Fund. Louise Black is funded by the National Institute for Health Research and Rosie Mansfield is funded by the Department for Education.

Correspondence concerning this article should be addressed to Louise Black, Manchester Institute of Education, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK. Email: louise.black@manchester.ac.uk

Abstract

The self-report version of the strengths and difficulties questionnaire (SDQ) is widely used in clinical and research settings. However, the measure's suitability for younger adolescents has recently been called into question by readability analysis. To provide further insight into the age-appropriateness of the SDQ, readability was assessed at the item level alongside consideration of item quality criteria, its factor structure was analyzed, and measurement invariance between adolescents in year seven (age 11–12) versus year nine (age 13–15) was tested. The measure showed a wide range of reading ages, and the theorized factor structure was unacceptable. Measurement invariance was therefore considered for a flexible exploratory structural equation model, and no evidence of differences between age groups was found. Suggestions are made for the measure's revision based on these findings.

Keywords: strengths and difficulties questionnaire; readability; measurement invariance; mental health; adolescents.

Age appropriateness of the self-report Strengths and Difficulties Questionnaire

The self-report version of the Strengths and Difficulties Questionnaire (SDQ) is a popular measure of mental health in 11–16-year-olds (Goodman et al., 1998; Johnston & Gowers, 2005) that has been extensively used in epidemiological research (e.g., Hafekost et al., 2016; NHS Digital, 2018; Polanczyk et al., 2015). Self-report measures are generally attractive in research, particularly in longitudinal and large-scale studies. This is partly because young people can be easier to recruit than parents, and data burden is reduced compared to teacher report methods (Humphrey & Wigelsworth, 2016). Moreover, such measures allow direct assessment of the young person's perspective in accordance with policy recommendations (Deighton et al., 2014). Despite these advantages, scale- and subscale-level analysis suggest the SDQ may be unsuitable for those with reading ages below 13–14 (Patalay et al., 2018). Not only is this higher than the intended 11-year-old population, it also exceeds general scale development recommendations, which suggest that measures should never exceed the reading level of a 12-year-old (Terwee et al., 2007). There is also evidence to suggest that the reading age of individuals can be up to 5 grades lower than their reported education grade, especially for those experiencing mental health difficulties (Jackson et al., 1991; Jensen et al., 2006). There is, therefore, a need for better understanding of the age appropriateness of this measure.

Though the self-report SDQ has been consistently employed in large national studies (e.g., Hafekost et al., 2016; NHS Digital, 2018), and has been recommended for research and clinical settings (Vostanis, 2006; Wolpert et al., 2015), robust evidence of its factor structure is scant. Two review articles have broadly advocated for the use of the self-report SDQ, as a well-validated measure (Vostanis, 2006; Wolpert et al., 2015). However, it should be noted that psychometric evidence underpinning their recommendations often related to translated versions, though psychometric characteristics are likely version dependent (Flake et al., 2017). Indeed, the self-report SDQ has shown only partial measurement invariance across different language versions (Ortuño-Sierra, Fonseca-Pedrero, et al., 2015).

Furthermore, studies of the English version on which recommendations were made particularly failed to report model fit (Goodman, 2001; Goodman et al., 1998). Though exploratory factor analysis was used in the original study, a 5-factor solution was retained despite substantial cross-loadings for seven items (Goodman, 2001), suggesting potential problems with the structure. Where confirmatory factor analysis (CFA) techniques were employed to analyze the self-report English version, the proposed structure was also shown to be problematic, with inconsistent fit based on recommended guidelines. These suggest values of around .95 for the comparative fit index (CFI) and around .06 for the root mean square error of approximation (RMSEA) can be judged to be acceptable (Hu & Bentler, 1999). Goodman et al. (2010) found CFI = .837 and RMSEA = .063 via weighted least squares means and variance adjusted (WLSMV), while Percy et al. (2008) reported CFI = .817 and RMSEA = .047 via robust maximum likelihood (MLR) estimation. The consistently low CFI may be due to problems with the pattern of covariances specified by the model, consistent with the known substantial cross-loadings (Goodman, 2001; Percy et al., 2008), though discrepancies between RMSEA and CFI can occur for many different reasons (see Lai & Green, 2016, for more details). The fact that both studies include adolescents as young as 11–12 may also have contributed to model misfit.

This lack of clear support for the self-report SDQ's factor structure suggests a need for more detailed examination of its psychometric qualities, as has been explicitly called for in a recent systematic review (Bentley et al., 2019). This is particularly necessary given the centrality of the measure in adolescent mental health research (e.g., Deighton et al., 2019; Dray et al., 2016; Hafekost et al., 2016; NHS Digital, 2018; Polanczyk et al., 2015; Wigelsworth et al., 2012). Although evidence based on the SDQ suggests an increase in mental health difficulties in mid adolescence, around ages 14–15 (Deighton et al., 2019; Dray et al., 2016), it is not clear whether differences between early adolescents, around ages 11–12, and the 14–15 age group are due to differences in measurement properties, or the SDQ's high reading age (Patalay et al., 2018). Indeed, measurement invariance between different age groups is yet to be examined, which we therefore sought to address in the current study. The choice of age groups in the current study was selected for pragmatic reasons since we conducted secondary data analysis. Nevertheless, the use of this dataset enabled examination of the key transition to mid adolescence. It also allowed comparison between the SDQ's youngest intended age (11 years old), as per its original validation (Goodman et al., 1998), and the recommended minimum age (13 years old) based on recent readability findings (Patalay et al., 2018).

While the analysis of readability by Patalay et al. (2018) provided valuable insight into the age appropriateness of the measure, readability was only considered for whole subscales meaning three issues remain unexplored. First, while considering items together as subscales or whole measures allows the use of texts of more appropriate length for readability formulas, information is lost about individual items (Oakland & Lane, 2004). Second, the presentation of items in accordance with psychometric best practice, including factor structure, should also be considered. For instance, items should have appropriate response formats and consist of single statements to avoid confusion (Saris, 2014; Terwee et al., 2007). Finally, while age invariance of the proxy version has been considered (He et al., 2013), measurement invariance of the self-report English instrument has not been tested, to our knowledge. Based on these identified gaps, we aimed to explore the following for the self-report SDQ: 1) item-level readability, 2) item quality, 3) the factor structure, and 4) age measurement invariance between English secondary school students in year seven (age 11–12) and year nine (age 13–15). We hypothesized the reading age to be higher than the intended population, consistent with Patalay et al. (2018) and that item quality would vary according to psychometric criteria (this has not been evaluated previously and was therefore exploratory). Given that findings on the structure of SDQ have been conflicting, we were unable to hypothesize which structure would be the most appropriate, thus the third aim of our study was also necessarily exploratory. Finally, we hypothesized non-measurement invariance between the two age groups, as we expected the year nine group to have a better understanding of the items, based on previous readability evidence (Patalay et al., 2018).

Method

Secondary data analysis was conducted of a large project aimed at promoting resilience in six areas of England, chosen on the basis of need. The original dataset consisted of 30,842 students, though 552 cases were excluded (1.8%) from current analyses since these had missing data for all SDQ items. Students were in year seven (50.7%, aged 11–12, *M* = 12.21, *SD* = 0.29) and nine (49.3%, aged 13–15, *M* = 14.20, *SD* = 0.29) from 114 schools (52.4% female). The ethnicity of our sample was very similar to national figures (Department for Education, 2017b) with 74.1% white, 9.5% Asian, 5.7% Black, 3.9 Mixed, .2% Chinese, 1.5% any other ethnic background, and 1.2% unclassified. The proportion of pupils with a special educational need was 11.6%, compared to the national figure of 14.4% (Department for

Education, 2017c). Rates of low income were above average in this community sample, given the focus of the project: The percentage of students who had ever been eligible for free school meals was 36.4% which was above the national average of 29.1% for those eligible in the previous six years (Department for Education, 2017a).

Total difficulties scores for the SDQ were also above rates expected in community samples, based on the measure's 20-year-old bandings (Goodman et al., 1998): 62.2% scored in the 'normal' range compared to 80% in the validation sample, 18.4% scored in the 'borderline' range compared to 10% in the validation sample and 19.6% scored in the 'abnormal' range compared to 10% in the validation sample. However, self-reported psychological wellbeing in the current sample (*M* = 23.88, *SD* = 5.33) was similar to the average found in a nationally representative sample of 16–24-year-olds (M = 23.57, SD = 3.61; Ng Fat et al., 2017). Reading ability was also below average based on end of primary school test results, with 63% of the year seven cohort reaching the expected grade compared to the national result of 66% (Department for Education, 2016), and 72.2% of the year nine cohort compared to the national result of 78% (Department for Education, 2014).

Following approval by the UCL Research Ethics Committee (UCL Ref: 8097/003) survey data were collected via a secure online portal during the normal school day from students whose parents had not opted out. The SDQ was completed as part of a battery of measures, all of which had explanations for items found to raise issues during piloting. These were constructed to help pupils without altering items, and since researchers did not administer the survey face-to-face they could not respond to queries. Pupils were instructed that these could be obtained by hovering their mouse over certain words. For example, if pupils hovered over the word "restless", they were given the explanation "unable to stay still".
All items which had explanations are indicated in Table 2.

Students responded to the 25-item SDQ (Goodman et al., 1998) using a 3-point Likert scale (*not true*, *somewhat true*, *certainly true*). These 25 items form five subscales of five items each (more detail on the content of items can be found in Table 2). Internal consistency coefficients are presented in several formats to reflect both the typically reported standard (Cronbach's alpha), as well as formulae that account for violations likely present in the data (see Table 1). Ordinal alpha accounts for the ordinal nature of Likert items since it is based on the polychoric correlation matrix (Gadermann et al., 2012), while

McDonald's omega is a model-based reliability which does not assume tau-equivalence (Raykov & Marcoulides, 2016). In line with other assessments of the SDQ (Bøe et al., 2016; Panayiotou et al., 2019), ordinal alpha and omega were shown to be higher than Cronbach's alpha in the current sample (see Table 1). ESEM factor loadings can also be found in supplemental table S1.

Table 1.

SDQ Subscale Reliability Coefficients

	Cronbach's α	Ordinal α	McDonald's ω [95% CI]
Emotional problems	.74	.81	.74 [.73, .74]
Conduct problems	.64	.76	.66 [.65, .66]
Hyperactivity	.74	.80	.75 [.74, .75]
Peer problems	.59	.72	.60 [.59, .61]
Prosocial	.69	.78	.69 [.68, .69]
Total difficulties	.81	.86	.87[h] [.87, .87]

Note. [h] Hierarchical omega coefficient

Analysis

Readability Testing

Calculating multiple readability estimates is recommended given the lack of a gold standard readability formula, and the variability in their focus (Janan & Wray, 2012). The current study applied four widely used and established readability assessments, all of which are calculated by incorporating different text components. The Dale-Chall Readability Formula (DC; Chall & Dale, 1995;
Dale & Chall, 1948), considers the percentage of difficult words, and the average sentence length.

Difficult words are those that do not appear on the Dale-Chall Readability word list:

$$DC = 0.1579(DW/TW \times 100) + 0.0496(AWS) + 3.6365$$

Where *DW = total number of difficult words, TW = total number of words, AWS = average number of words per sentence.*

The Flesch-Kincaid Reading Grade (FK; Kincaid et al., 1975), considers average syllables per word and the average sentence length:

$$FK = (0.39 \times AWS) + (11.8 \times ASW) - 15.59$$

where *AWS = average number of words per sentence; ASW = average number syllables per word.* The Gunning Fog Index (GFI; Gunning, 1952) considers number of words, sentences and hard words (those with three syllables or more):

$$GFI = 0.4 \times [\ AWS + (100HW / TW)]$$

where *AWS = average number of words per sentence; HW = total number of hard words; TW = total number of words.*

Finally, the Coleman Liau Index (CLI; Coleman & Liau, 1975) incorporates number of letters instead of syllables:

$$CLI = (0.0588 \times LW) - (0.296 \times SW) - 15.8$$

where *LW = average number of letters per100 words; SW = average number of sentences per 100 words.*

All indices provide readability as a US grade-level. The readability of SDQ items and subscales was then calculated by averaging the US-grade level score of the four indices, and then adding six to get the average reading age. The age appropriateness of SDQ items was judged against the original minimum recommended age of 11 (Goodman et al., 1998).

Item Quality Criteria

Consistent with readability indices, psychometric guidance suggests scale items should be simple in language and grammar, regardless of the age of the target population (Irwing & Hughes, 2018; Terwee et al., 2007). Beyond this, other important aspects of the content and structure of items must be considered alongside readability tests, for a more comprehensive assessment (Oakland & Lane, 2004). Additional item quality criteria deemed relevant to age-appropriateness and mental health were therefore identified to supplement readability analyses. First, items should ideally consist of single statements

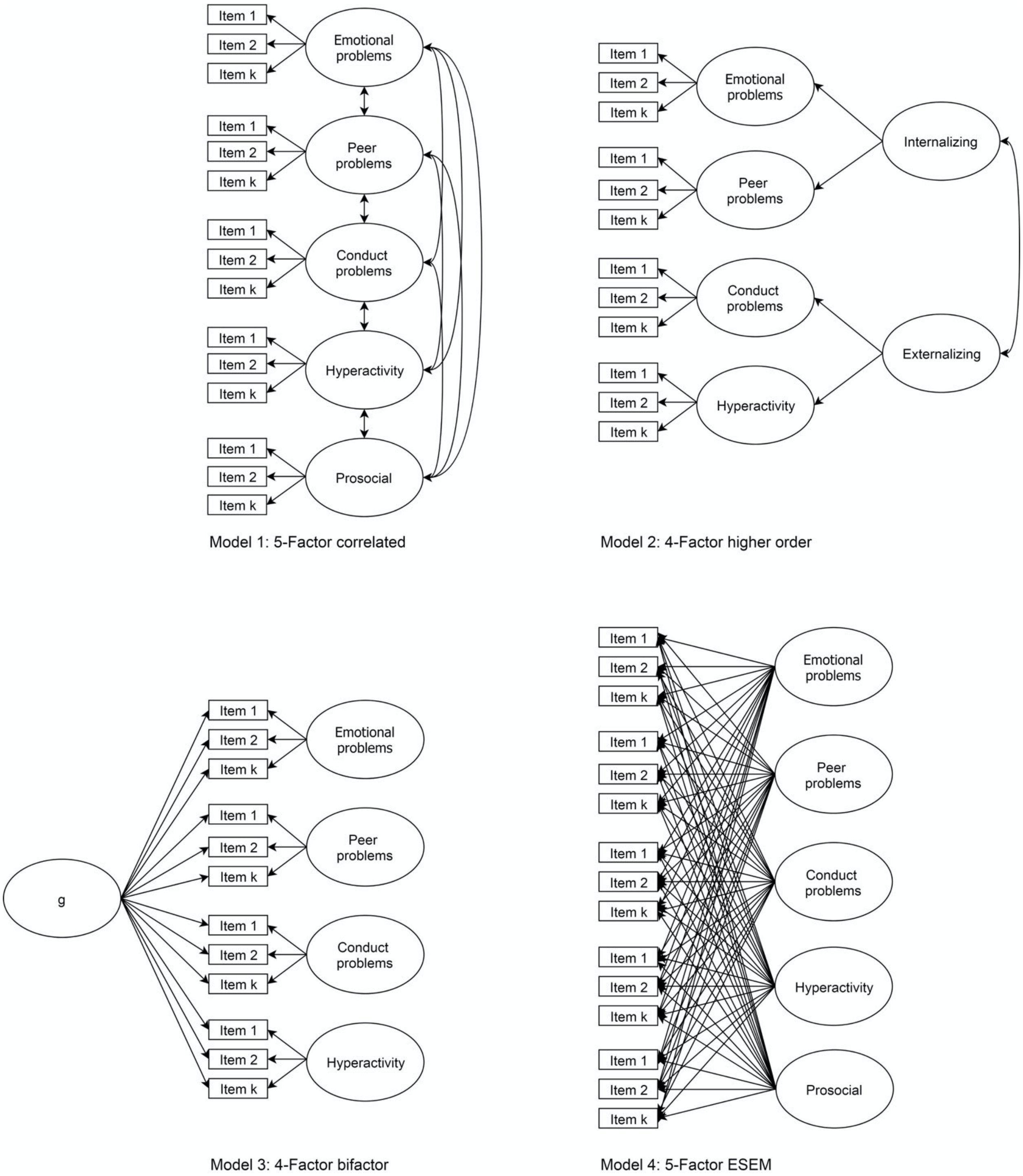
(Irwing & Hughes, 2018; Saris, 2014; Terwee et al., 2007), and avoid reverse wording to reduce confusion (Irwing & Hughes, 2018; van Sonderen et al., 2013). Floor and ceiling effects (endorsement of the lowest or highest response at > 15%) should not be present. Absence of these is an indication that measures reliably distinguish individuals across the range of symptoms (Terwee et al., 2007). Items should also be presented with a clear and appropriate reference period to the concept under study (Irwing & Hughes, 2018; Saris, 2014). Since all items had the same reference period, we used the first three criteria to assess items and considered those that satisfied two out of three to be of higher quality.

Factor Structure and Measurement Invariance

Given the poor factor structure of the self-report SDQ in other samples (Goodman et al., 2010; Goodman, 2001; Percy et al., 2008), we considered both CFA and ESEM with geomin rotation (see Figure 1). We estimated three CFA models, the first of which was a correlated structure of the five subscales, based on the original theoretical structure of the measure representing the five subscales typically used (Goodman, 2001). Secondly, we included a correlated 2factor higher-order structure in which emotional problems and peer problems loaded onto a second-order internalizing factor, and conduct problems and hyperactivity loaded onto a second-order externalizing factor as suggested elsewhere (Goodman et al., 2010). Thirdly, we estimated a bifactor model (Chen & Zhang, 2018) with a general difficulties factor, and four residual difficulty subdomain factors. This model has shown some promise in other language versions (e.g., Ortuño-Sierra, Chocarro, et al., 2015) and allows the total difficulties subscale to be represented as a general factor after accounting for specific variance captured by each of the four problem domains. The prosocial factor was excluded from both the bifactor and higher-order models since these were used to examine the hypothesized 4-factor total difficulties score (Goodman, 2001). We finally tested a 5-factor ESEM model, which was used to explore age measurement invariance.

Figure 1

Models Tested



Model 1: 5-Factor correlated

Model 2: 4-Factor higher order

Model 3: 4-Factor bifactor

Model 4: 5-Factor ESEM

Where measures lack proposed dimensionality, as is the case with the self-report SDQ (Goodman et al., 2010; Goodman, 2001; Percy et al., 2008), and invariance testing is warranted, given recent claims about age (Deighton et al., 2019; Dray et al., 2016; NHS Digital, 2018), exploratory structural equation modeling (ESEM) techniques can be used (Marsh, Nagengast, et al., 2013). As others have pointed out, though ESEM structures should not be used to conceal problems with a measure, they can provide a more realistic framework for measurement invariance analysis where CFA models do not fit sufficiently well (Tóth-Király et al., 2017). Furthermore, given the substantial cross-loadings and shared variance in the SDQ (Goodman, 2001; Percy et al., 2008), ESEM can provide a more robust approach than post-hoc addition of parameters (e.g., crossloadings) following modification indices (Chiorri et al., 2016). We therefore opted to extract five factors in line with the original theoretical model, but in ESEM every item is permitted to load onto every factor so that shared variance in the data is not misspecified.

When accounting for the fact data were sampled from pupils clustered in schools (using type = complex), the ESEM models required greater numbers of parameters to be estimated than there were schools in the sample (165 > 114), thus resulting in a warning about the trustworthiness of standard errors. Given that the implications of this in model estimation are not well understood (Muthén & Muthén, 2016), and parameter estimates would not be directly affected, clustering effects were not controlled for. This decision was guided by the small intra-cluster correlations for the SDQ variables (<.05) and the fact that controlling for clustering made little difference to the standard errors and therefore conclusions (results can be provided upon request). For consistency we therefore did not account for clustering in any model.

Chi-square difference testing is typically used to compare the fit of measurement invariance models. However, its sensitivity to sample size made this inappropriate for our study, suggesting approximate fit indices should be used. Since the majority of measurement invariance simulations focusing on performance of fit indices have treated items as continuous (Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008), the degree to which common fit indices are appropriate for comparing models using polychoric matrices and WLSMV is unclear. For instance, given that the chis-quare of WLSMV is not comparable in the same way as for maximum likelihood, CFI comparisons might not be appropriate in these cases (Sass et al., 2014). Analyses were therefore conducted in Mplus 8.3

using MLR and treating items as continuous. This also allowed us to account for the non-normality of the data and enabled missing data to be handled via full information maximum likelihood under the assumption of missing at random (Muthén & Muthén, 1998-2017). All cases with data for at least one SDQ item were therefore included in our analysis. Though items were treated as continuous, floor effects were likely in a screening measure, so sensitivity tests for the CFA and ESEM models were conducted, in which items were treated as ordinal using WLSMV (Brown, 2015; Li, 2016).[4]

Model fit was judged in line with published recommendations. Chi-square statistics are reported, but not interpreted as indicating fit given their known sensitivity to sample size. The CFI and the Tucker Lewis index (TLI) were considered to be acceptable at around .95, and RMSEA around .06 (Hu & Bentler, 1999). The standardized root mean squared residual (SRMR) was considered to be acceptable < .08 in the absence of any large residuals (Asparouhov & Muthén, 2018). In addition to these standardized indices, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are also reported to compare models with the same outcome variables, with lower values indicating better model fit.

Measurement invariance testing was conducted by estimating baseline models for each age group separately, followed by a configural model in which parameters were freely estimated in each group, a metric model with loadings constrained to be equal across groups, and finally, a scalar model in which intercepts were also held equal (Muthén & Muthén, 1998-2017). Given the large sample size, CFI difference (DCFI) was used to judge approximate invariance (Sass et al., 2014). In line with wider ESEM literature (Marsh, Nagengast, et al., 2013; Marsh, Vallerand, et al., 2013; Tóth-Király et al., 2017), and specific invariance analysis of the SDQ (Chiorri et al., 2016), we adopted a threshold of .01 for DCFI. This cutoff has been shown to perform well with the Mplus calculation of CFI and under different conditions of invariance and non-invariance (Chen, 2007).

Results

[4] WLSMV solutions were not estimated for measurement invariance testing, given the problems with comparing CFI for this estimator (Sass et al., 2014).

Readability Estimates

Table 2 presents the four readability estimates by US grade-level, the average across the four indices, and the reading age in years. Estimates were calculated for the introductory text, individual items, subscales and total scale. The introductory text was found to have a reading age considerably greater than 11. Similarly, items 3, 13, 16 (emotional), 4, 20 (prosocial), 10, 15 (hyperactivity), and 14 (peer problems) were calculated as having readability estimates greater than 12 years old. Of the five subscales, emotional problems and hyperactivity were calculated as having the highest reading ages (>12). However, despite appropriate estimates for the remaining subscales and total scale, conduct problems was the only subscale not to include any items with a reading age greater than 12 years. Items 10, 13, 15, 16 and 20 were of particular concern with reading ages greater than 15 years.

Item Quality Criteria

The measure's items, floor/ceiling effects, and quality scores can be found in Table 2. While we expected varied quality, results were not favorable with 17 items (68%) shown to have poor item quality (see Table 2). Specifically, of the SDQ's 25 items, 14 (four emotional problems, four conduct problems, three hyperactivity-inattention, two prosocial, and one peer problems) clearly include more than one statement, and therefore request a response about more than one experience. The measure also has five reversed items across the conduct problems, hyperactivity, and peer problems scales. All 20 difficulties items showed substantial floor effects, ranging from 21–85%, and a further eight also had ceiling effects, ranging from 15–34 %. The prosocial items showed ceiling effects, ranging from 29-69%, and one also had a floor effect at 16%.

Table 2.

SDQ Items Floor/Ceiling Effects, Readability Estimates by US Grade-Level, Average Estimate Across Indices and Reading Age

Floor–Ceiling %	Score	Instructions and Items	US-Grade Level				Average US Grade Level	Age
			DC	FK	GFI	CLI		
		For each item, please mark the box for Not True, Somewhat True or Certainly True. It would help us if you answered all items as best you can even if you are not absolutely certain or the item seems daft! Please give your answers on the basis of how things have been for you over the last six months.	6.49	6.88	9.22	6.21	7.20	13.20†
		Emotional problems	**6.86**	**5.52**	**9.90**	**4.43**	**6.68**	**12.68†**
43.7–18.4	-	3) I get a lot of headaches, stomach-aches or sickness	5.84	4.91	8.04	7.69	6.62	12.62†
31.1–29.6	+	8) I worry a lot	3.83	0.72	1.60	-8.51	-0.59	5.41
55.6–11.7	-	13) I am often unhappy, down-hearted or tearful	6.24	9.09	14.23	9.36	9.73	15.73†

27.3–30.7	-	16) I am <u>nervous</u> in new situations. I easily lose confidence	10.20	7.60	14.00	5.32	9.28	15.28†
50.9–14.6	-	24) I have many fears, I am easily scared	6.01	3.81	8.20	1.81	4.96	10.96
		Conduct problems	**4.69**	**2.82**	**4.08**	**1.45**	**3.26**	**9.26**
39.4–23	-	5) I get very angry and often lose my temper	4.08	4.91	3.60	2.47	3.77	9.77
39.5–7.4	-	7) I usually do as I am told (R)	6.24	4.01	8.51	-4.08	3.67	9.67
74.7–5	-	12) I fight a lot. I can make other people do what I want	3.96	0.52	2.60	-2.27	1.20	7.20
56.6–14	-	18) I am often <u>accused of</u> lying or cheating	6.01	5.23	3.20	4.01	4.61	10.61
82.7–4	+	22) I take things that are not mine from home, school or elsewhere	4.23	1.83	4.80	6.23	4.27	10.27
		Hyperactivity	**7.06**	**4.66**	**9.56**	**4.96**	**6.56**	**12.56†**
21.1–34.1	-	2) I am <u>restless</u>, I cannot stay still for long	5.84	2.32	3.60	3.12	3.72	9.72
35.7–27	+	10) I am constantly <u>fidgeting</u> or squirming	11.83	8.34	15.73	11.60	11.88	17.88†

26.3–30.6	-	15) I am easily distracted, I find it difficult to concentrate	10.45	9.55	20.00	9.46	12.36	18.36†
28.7–15.1	-	21) I think before I do things (R)	3.93	0.56	2.40	-0.16	1.68	7.68
31.4–13	-	25) I finish the work I'm doing. My attention is good (R)	3.88	2.88	6.00	1.20	3.49	9.49
		Peer problems	**4.99**	**4.13**	**5.75**	**2.85**	**4.43**	**10.43**
63.8–10	-	6) I am usually on my own. I generally play alone or keep to myself	5.11	5.67	8.51	0.96	5.06	11.06
85.1–2.8	-	11) I have one good friend or more (R)	3.98	1.06	2.80	0.12	1.46	7.46
40–9	-	14) Other people my age generally like me (R)	3.98	7.32	8.51	6.00	6.46	12.46†
70.3–8.2	+	19) Other children or young people pick on me or bully me	5.62	4.75	4.40	4.49	4.81	10.81
44.5–14	+	23) I get on better with adults than with people my own age	5.55	3.84	4.80	3.29	4.37	10.37
		Prosocial	**5.35**	**5.23**	**4.87**	**5.36**	**5.20**	**11.20**
2.9–59.8	-	1) I try to be nice to other people. I care about their feelings	5.17	2.40	2.60	1.35	2.88	8.88

8.6–47.6	+	4) I usually share with others (food, games, pens etc.)	7.59	8.92	8.04	6.39	7.74	13.74†
5.6–51	-	9) I am helpful if someone is hurt, upset or feeling ill	4.18	4.75	4.40	3.96	4.32	10.32
4.9–69	+	17) I am kind to younger children	3.93	2.44	2.40	2.78	2.89	8.89
16.4–28.7	+	20) I often <u>volunteer</u> to help others (parents, teachers, children)	5.84	8.92	8.04	13.57	9.09	15.09†
		Total Scale (without instructions)	**5.68**	**4.35**	**6.55**	**3.72**	**5.08**	**11.08**
		Total Scale (with instructions)	**5.78**	**4.31**	**6.54**	**4.24**	**5.22**	**11.22**

Note. In bold are the estimates for the subscales and total scale. Underlined words are those for which additional explanations were provided when the mouse was hovered over them in the online administration. + = high quality; - = low quality. (R) = Reversed items. **† =** items and scales with readability age above 11. DC = Dale-Chall Readability Formula; FK = Flesch-Kincaid Reading Grade; GFI = Gunning Fog Index; CLI = Coleman Liau Index.

Readability vs. Item Quality

Though our readability methodology suffers from applying formulas to short texts (Oakland & Lane, 2004), this was considered alongside item quality criteria, so that items could be evaluated more comprehensively. For instance, the item with the lowest reading age, "*I worry a lot*", also performed well in terms of item quality since it is not reversed, and consists of a single statement. Conversely, the item "*I fight a lot. I can make other people do what I want*" has a low reading age, but introduces confusion since respondents must affirm two independent behaviors. Another consideration is that the measure is often deployed in schools, as was the case for our sample (e.g., Wigelsworth et al., 2012). The item "*I am easily distracted, I find it difficult to concentrate*" has the highest reading age because it contains several multiple syllable words. On one hand, young people in schools may regularly be talked to about concentration and therefore be more readily primed to recognize these words than readability formulas would suggest. However, item quality criteria confirm that this statement is unnecessarily complex, containing two statements. Readability and age-appropriateness of measures are therefore more complex than any one type of analysis might suggest.

Factor Structure and Measurement Invariance

School year group was available for all but one participant, and missingness for SDQ responses ranged from .5-1.5%. Variance and Covariance coverage were high (>.97) for SDQ items suggesting that estimates were likely to be trustworthy (Muthén et al., 2017). Since data were not missing completely at random, χ^2 (13289)= 17509.62, $p < .0001$, we explored missingness at the subscale level, using gender, age, ethnicity, self-reported wellbeing, special educational needs and free school meal eligibility as predictors. Special educational needs (OR = .25–.37) predicted less missing data for all subscales. Unclassified ethnicity predicted less missing data for all but the conduct problems subscale (OR = .01–.21). Asian ethnicity predicted less missing data for peer problems, prosocial behaviour and hyperactivity (OR = .28–.35). Higher wellbeing predicted less missing data for peer problems and prosocial behaviour (OR = .92–.93), while girls (OR = .33) and those from black ethnic backgrounds (OR = .28) were less likely to have missing data for prosocial behaviour.

Fit of all models estimated is provided in Table 3. The original correlated 5-Factor structure was found to have poor fit, as did the higher-order model. The bifactor structure of the four difficulties subscales similarly indicated a total difficulties score to be problematic, even though bifactor structures are highly parameterized with a tendency to overfit (Murray & Johnson, 2013). As expected, given the flexibility of such models, the ESEM solution provided a much better fit to the data. Nevertheless, primary ESEM loadings were strongly related to their corresponding parameters in the CFA model. This was established via a correlation between loadings from the ESEM and CFA models (*r* = .65) following the example by Marsh, Vallerand, et al. (2013).

Table 3.

Model Fit for Main and Sensitivity Analysis Models

Model	Estimator	χ^2 (df)	AIC	BIC	RMSEA [90% CI]	CFI	TLI	SRMR	λ	h^2
5-Factor correlated	MLR	27966.58 (265)**	1357768.55	1358475.63	.059 [.058, .059]	.807	.781	.063	.351–.716	.123–.513
	WLSMV	44612.71 (265)**	-	-	.074 [.074, .075]	.832	.810	.081	.480–.843	.221–.710
4-Factor higher-order	MLR	19841.26 (165)**	1111239.31	1111779.98	.063 [.062, .064]	.822	.795	.057	.323–.967	.104–.488
	WLSMV	28588.03 (165)**	-	-	.075 [.075, .076]	.867	.847	.071	.461–.983	.213–.726
4-Factor bifactor	MLR	19087.85 (150)**	1109992.68	1110658.12	.065 [.064, .065]	.829	.783	.069	-.086–.707	.146–.635
	WLSMV	36973.36 (150)**	-	-	.090 [.089, .091]	.828	.782	.080	.171–.884	.206–.666
ESEM	MLR	5791.31 (185)**	1333225.12	1334597.69	.032 [.031, .032]	.961	.937	.016	-	.189–.558

	WLSMV	6514.99 (185)**	-	-	.034 [.033, .034]	.976	.961	.016	-	.270–.694

Note. ESEM = exploratory structural equation modeling; MLR = robust maximum likelihood; WLSMV = weighted least square mean and variance adjusted; AIC = Akaike information criterion; BIC = Bayesian information criterion; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis Index; SRMR = standardized root mean squared residual; λ = factor loadings; h^2 = item communalities.

***p* < .01.

The ESEM solution (see supplemental table S1) revealed nine items to cross-load with a discrepancy of < .30 between the highest and second highest loadings, which is indicative of problems with the item (Matsunaga, 2010). Each of the five reversed items also loaded above .34 on the prosocial factor, and less strongly on their theorized difficulties factors. The prosocial factor was not correlated with the emotional problems and peer problems factors at a significant level. Similarly, the hyperactivity factor was not significantly associated with the peer problems factor. Factor correlations beyond this were in expected directions, with the largest associations seen between hyperactivity and conduct problems (*r* = .49), and emotional problems and peer problems (*r* = .38). Sensitivity analysis also revealed that accounting for the categorical nature of items via WLSMV had little impact on results. No changes in fit or loadings were seen in terms of recommended cutoffs, supporting confidence in the main results reported based on MLR.

Acceptable model fit was found for the two age groups separately. Consistent with findings for the parent version with middle and older adolescents (He et al., 2013), but counter to our hypothesis based on previous readability evidence, approximate age measurement invariance was supported, as the DCFI was found to be below .01 in all comparisons (see Table 4).

Table 4.

ESEM Age Measurement Invariance Findings

Model	χ^2 (df)	AIC	BIC	RMSEA [90% CI]	CFI	TLI	SRMR	**Δ**χ^2 (df)	DCFI	h_2
Y7 Baseline	2779.44 (185)**	676262.55	677525.25	.030 [.029, .031]	.964	.941	.016			.194–.572
Y9 Baseline	3191.19 (185)**	653776.06	655029.61	.033 [.032, .034]	.957	.931	.018			.186–.557
Configural	5967.77 (370)**	1330038.61	1332783.73	.032 [.031, .032]	.961	.936	.017			
vs. Metric	6280.37 (470)**	1330320.63	1332233.90	.029 [.028, .029]	.959	.948	.020	409.27 (100)**	.002	
vs. Scalar	6920.98 (490)**	1330917.70	1332664.59	.029 [.029, .030]	.955	.945	.021	729.63 (20) **	.004	

Note. Robust maximum likelihood was used. ESEM = exploratory structural equation modeling; AIC = Akaike information criterion; BIC = Bayesian information criterion; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis Index; SRMR = standardized root mean squared residual; **Δ**χ^2 = chi-square difference test; DCFI = CFI difference; h^2 = item communalities.

**p < .01.

Discussion

Though the self-report SDQ is widely used, including to study age differences (Deighton et al., 2019; Hafekost et al., 2016; Johnston & Gowers, 2005), evidence of its age appropriateness has been limited. Building on existing evidence (He et al., 2013; Patalay et al., 2018) we addressed this gap by considering the measure's item-level readability, item quality, factor structure, and age measurement invariance. Items showed a wide range of reading ages, which was more varied than previous subscale-level analysis had indicated (Patalay et al., 2018). Many items also appeared to be too difficult for the intended age group. Beyond this, a substantial proportion of the measure was found to be problematic in terms of item quality, and the proposed factor structure was a poor fit to the data. ESEM allowed approximate measurement invariance to be tested between students in year seven versus year nine, which suggested that this flexible structure was invariant across these groups.

While Patalay and colleagues (2018) had already demonstrated the measure may not be suitable for adolescents under 13, their analysis was unable to clarify which items might be problematic. In fact, our results suggest scale and subscale-level reading scores could be misleading since they suggested levels around age 11. Counter to our first hypothesis, item-level readability was much more varied than that found previously at the subscale level. We found some items to be much more difficult and others much easier. For instance, while the emotional problems subscale had an average reading age of 12.68, the item "*I worry a lot*" performed much better with an average reading age of 5.41. This item is therefore an example of optimal simplicity.

Beyond the item-level analysis, the instructions did not meet recommendations published elsewhere that even adult scales should have reading ages of no more than 12 (Terwee et al., 2007). This suggests there may have been problems even for higher quality items. In fact, special attention to instructions has been recommended for surveys with young people since clearer and more detailed instructions can be associated with greater reliability (Omrani et al., 2018). Similarly, Though the stated reference period in the SDQ instructions is clear, i.e. not subjective such as "often", but finite, "over the last six months", this may not be appropriate to the assessment of symptoms in adolescents. Younger adolescents, in particular, tend to find long reference periods challenging, and guidelines suggest very recent or current reference periods may lead to more valid responses in this age group (Bell, 2007; de Leeuw, 2011).

As well as clarifying readability analysis, consideration of item quality criteria revealed the measure to have certain other problems. Alongside the fact that over half of items contain multiple statements, the SDQ also contains five reversed items. While such items are common in scale development, it is generally advised that these be avoided since they tend not to factor well with other constructs or be opposite indicators as developers intend them to be (Ebesutani et al., 2012; Suárez-Alvarez et al., 2018; van Sonderen et al., 2013). In the current study it was clear the reversed items were not measuring the subscale constructs cleanly, as ESEM results revealed all these items to have substantial cross-loadings. This is also consistent with findings in other language versions of the SDQ (Garrido et al., 2018; van de Looij-Jansen et al., 2011). Specifically, we found each of the reversed items loaded more strongly on the prosocial factor than on their respective theorized factors. Some shared variance could reasonably be anticipated. However, the magnitude of these cross loadings (particularly on the prosocial factor), suggests that beyond age-appropriateness, these items may also face wider validity problems. Reversed items can affect instrument structure through misresponse since their content may not be perceived as opposite to positively worded statements (Weijters & Baumgartner, 2012). Though we did not explicitly examine common method effects, our ESEM results suggest reversed items could have introduced noise into the structure through similarity to prosocial items, as they all relate to positive behaviors.

Item quality criteria also provided insight into the measure’s applicability across the range of symptoms. In our community sample, which showed above average levels of mental health difficulties, high levels of floor or ceiling effects were seen for every item. While this is a common feature of clinical measures used in samples with predominantly healthy individuals, the measure’s use may be somewhat limited, particularly if recommended dimensional approaches to understanding symptoms are adopted (Krueger et al., 2018). This is because measures with high floor and ceiling effects tend to have less discriminatory ability and responsiveness; in other words they may be less able to detect change and discriminate between individuals with different levels of problems (e.g., high versus borderline; de Vet et al., 2011). The three-point response format may contribute to the skewed nature of the data since having more categories can be associated with higher reliability and validity (Lozano et al., 2008). While there is relatively little research on number of response categories with young people, available evidence suggests around four options may provide a good balance in terms of memory, reading, reliability and stability (Bell, 2007; Omrani et al., 2018).

Beyond the issues already identified, further elements have also been suggested as indicators of psychometric quality. Of particular relevance to the current study, is that measures should ideally be developed in consultation with the target population (Irwing & Hughes, 2018; Terwee et al., 2007), since this allows assessment of acceptability and bias of items. It is possible that some of the psychometric problems identified in the SDQ are compounded by such issues, as to the authors' knowledge, such consultation did not take place in the development of the SDQ.

Regarding the SDQ's structure, we found the five correlated subdomains to be a poor fit to the data, and uncovered substantial shared variance across factors in the ESEM solution. Both the higher-order internalizing/externalizing model, and the bifactor difficulties model also failed to show good fit. These results indicate that using the SDQ to calculate subdomain scores is questionable (Raykov & Marcoulides, 2011). Our ESEM results further suggest the hypothesized structure may be problematic since several items loaded onto more than one factor.

The instrument's poor fit may also be explained by satisficing theory, which is considered to be of particular relevance to adolescents (Krosnick, 1991; Omrani et al., 2018). This holds that the greater the cognitive demand on participants, the lower the reliability of their responses, as steps involved in providing appropriate responses are skipped (Krosnick, 1991; Omrani et al., 2018). The following results in this study could support such an account: 1) subscales showed mixed reliability, as measured through internal consistency; 2) the instructions had a higher reading age than the lowest limit of the intended population; 3) many items did not have appropriate reading ages, with some at very high levels; 4) the reference period of six months is often considered to be inappropriate for younger adolescents (Bell, 2007; de Leeuw, 2011); 5) several items, particularly those with reverse wording, were found to tap into more than one construct; 6) many items contained multiple statements which tend to increase cognitive load (Oakland & Lane, 2004).

Since we found the hypothesized CFA structures to be inadequate, we proceeded to invariance testing with the ESEM model, which as expected showed excellent fit. We found no evidence of differences in how 11–12-year-olds versus 13–15-year-olds responded using this flexible model. Since we used DCFI to establish approximate invariance, we interpret our findings as suggesting that any differences between groups are likely insubstantial. Though we anticipated older students might respond markedly differently, as previous research suggested the SDQ may be more appropriate to their reading ability, (Patalay et al., 2018), our results suggest that both groups responded to it with the same level of ease and/or difficulty. Still, our readability evidence suggests that items with a reading age above 14 may have been too difficult for both groups. In fact our sample had below average ability in reading which could also support the idea that approximate invariance was caused in part by high reading age items being equally difficult for both groups. Further work is needed (e.g. cognitive interviews with young people) to consolidate our findings.

Taken together, our findings indicate a large proportion of self-report SDQ items are less appropriate for use with younger populations. The current study is the first to provide a detailed item-level readability analysis, thus uncovering specific issues with the self-report SDQ. While previous evidence suggested four of the five subscales had reading ages higher than the recommended minimum age (Patalay et al., 2018), the current study indicates this may be not be the case for *all* items. Still, our findings call for caution when using the self-report SDQ with younger adolescents or populations with mental health difficulties, since this group may have below average reading ability (Jensen et al., 2006; Moilanen et al., 2010). It should also be noted that self-report adolescent mental health measures have generally been found to be poor in terms of psychometric quality (Bentley et al., 2019). It is therefore important that researchers and clinicians consider carefully the psychometric quality and reading age of their chosen instrument in relation to their sample (Jensen et al., 2006).

Our study brought together robust and complementary methodological approaches to comprehensively assess age-appropriateness of a widely used measure for the first time. Indeed, our findings highlight the importance of conducting supplementary analysis such as readability and item quality alongside invariance testing, since these can provide additional insight. Together, assessment of item quality and readability with factor analysis suggested that the scale contains several difficult statements and psychometrically poor items with a response scale that prevents it from capturing the full spectrum of symptoms experienced in the general population (Terwee et al., 2007).

Despite these methodological strengths, a number of limitations must be acknowledged. First, though we attempted to overcome the problem of losing information about items when applying readability formulas to subscales, our item-level readability results should be interpreted carefully. These formulas were not designed for this purpose and therefore may not be as reliable as when used with longer passages (Oakland & Lane, 2004). However, we are confident that high-scoring items are likely inappropriate for younger audiences since they also showed poor item quality. It has also been suggested that assessment of readability at the item level is vital since this reflects how respondents actually perceive scale texts, particularly since individual items may be skipped or invalid responses provided when demands are too great (Calderón et al., 2006). In addition, although readability results were considered alongside other well-established indicators of item quality, these were not based on a standardized measure.

We also treated items as continuous so we could employ the more robust DCFI index for invariance testing, though our data were ordinal. The skewness in our data was controlled for by using MLR and sensitivity analysis using WLSMV supported these findings. Thirdly, though our large sample size was likely an asset for assessing the generalizability of floor and ceiling effects, and the factor structure of the measure, it is not currently clear how approximate difference testing using DCFI is affected by samples of the magnitude reported here. It is also possible that the explanations provided via the online portal affected measurement invariance by masking the differences in ability between the older and younger cohort. However, in any large-scale research with young people it is likely that support would be provided in some form (e.g. by a teacher or researcher). It is therefore likely very difficult to provide measurement invariance analysis across age groups without some kind of confound for ability.

Results must also be interpreted only for the ESEM model, which is less restrictive, with cross-loadings freely estimated. The theorized CFA model by Goodman et al. (1989) was not suitable for measurement invariance testing, and we therefore stress that invariance of this model could not be determined. Though lack of control over a priori structure in ESEM is therefore a limitation (Marsh et al., 2011), five factors corresponding to the original theoretical model were extracted in order to accommodate issues such as cross-loadings without resorting to post-hoc model modification. Similarly, though the large number of parameters in ESEM is a limitation, our large sample size was likely able to handle this with a ratio of 163.7 cases per parameter. Finally, though our sample was large, it was not representative of the general population since deprivation was seen at higher levels, given the focus of the project from which data were drawn.

Conclusion and Future Directions

While the self-report SDQ has been used extensively, our study suggests the measure would benefit from revisions two decades on from its original development. It is perhaps surprising that such a widely used measure suffers from issues such as those described here, although as our findings suggest, this is possibly due to the lack of attention to robust scale development practices (e.g. omission of cognitive interviews with young people). Items should be simplified, with reversed wording and multiple statements replaced with simpler alternatives, and more straightforward language used for items with high reading ages. We also recommend that such amendments be made in consultation with young people in line with policy and psychometric best practice (Deighton et al., 2014; Irwing & Hughes, 2018; Terwee et al., 2007).

References

Asparouhov, T., & Muthén, B. (2018). *SRMR in Mplus* http://www.statmodel.com/download/SRMR2.pdf

Bell, A. (2007). Designing and testing questionnaires for children. *Journal of Research in Nursing*, *12*(5), 461-469. https://doi.org/10.1177/1744987107079616

Bentley, N., Hartley, S., & Bucci, S. (2019). Systematic Review of Self-Report Measures of General Mental Health and Wellbeing in Adolescent Mental Health. *Clinical Child and Family Psychology Review*, *22*(2), 225-252. https://doi.org/10.1007/s10567-018-00273-x

Bøe, T., Hysing, M., Skogen, J. C., & Breivik, K. (2016). The Strengths and Difficulties Questionnaire (SDQ): Factor Structure and Gender Equivalence in Norwegian Adolescents. *PLOS ONE*, *11*(5), e0152202. https://doi.org/10.1371/journal.pone.0152202

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.

Calderón, J. L., Morales, L. S., Liu, H., & Hays, R. D. (2006). Variation in the readability of items within surveys. *American journal of medical quality : the official journal of the American College of Medical Quality*, *21*(1), 49-56. https://doi.org/10.1177/1062860605283572

Chall, J. S., & Dale, E. (1995). *Readability revisited : the new Dale-Chall readability formula.* Brookline Books.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Chen, F. F., & Zhang, Z. (2018). Bifactor Models in Psychometric Test Development. In *The Wiley Handbook of Psychometric Testing*. https://doi.org/doi:10.1002/9781118489772.ch12

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Chiorri, C., Hall, J., Casely-Hayford, J., & Malmberg, L.-E. (2016). Evaluating Measurement Invariance Between Parents Using the Strengths and Difficulties Questionnaire (SDQ). *23*(1), 63-74. https://doi.org/10.1177/1073191114568301

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283-284. https://doi.org/10.1037/h0076540

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability: Instructions. *Educational Research Bulletin*, *27*(2), 37-54. http://www.jstor.org/stable/1473669

de Leeuw, E. D. (2011). *Improving data quality when surveying children and adolescents: Cognitive and social development and its role in questionnaire construction and pretesting.* http://www.aka.fi/globalassets/awanhat/documents/tiedostot/lapset/presentations-of-theannual-seminar-10-12-may-2011/surveying-children-and-adolescents_de-leeuw.pdf

de Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: a practical guide*. Cambridge University Press.

Deighton, J., Croudace, T., Fonagy, P., Brown, J., Patalay, P., & Wolpert, M. (2014). Measuring mental health and wellbeing outcomes for children and adolescents to inform practice and policy: a review of child self-report measures. *Child and Adolescent Psychiatry and Mental Health*, *8*(1), 14. https://doi.org/10.1186/1753-2000-8-14

Deighton, J., Lereya, S. T., Casey, P., Patalay, P., Humphrey, N., & Wolpert, M. (2019). Prevalence of mental health problems in schools: poverty and other risk factors among 28 000 adolescents in England. *The British Journal of Psychiatry*, 1-3. https://doi.org/10.1192/bjp.2019.19

Department for Education. (2014). *Statistical First Release National curriculum assessments at key stage 2 in England, 2014 (Revised)*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/428838/SFR50_2014_Text.pdf

Department for Education. (2016). *National curriculum assessments at key stage 2 in England, 2016 (revised)*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/577296/SFR62_2016_text.pdf

Department for Education. (2017a). *Pupil premium: allocations and conditions of grant 2016 to 2017*. Retrieved 20/09/2018 from https://www.gov.uk/government/publications/pupil-premiumconditions-of-grant-2016-to-2017

Department for Education. (2017b). *Schools, pupils and their characteristics: January 2017*. Retrieved 15/05/2018 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650547/SFR28_2017_Main_Text.pdf

Department for Education. (2017c). *Special educational needs in England: January 2017*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633031/SFR37_2017_Main_Text.pdf

Dray, J., Bowman, J., Freund, M., Campbell, E., Hodder, R. K., Lecathelinais, C., & Wiggers, J. (2016). Mental health problems in a regional population of Australian adolescents: association with socio-demographic characteristics. *Child and Adolescent Psychiatry and Mental Health*, *10*(1), 32. https://doi.org/10.1186/s13034-016-0120-9

Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T. L., Damon, J. D., & Young, J. (2012). The Loneliness Questionnaire–Short Version: An Evaluation of Reverse-Worded and Non-Reverse-Worded Items Via Item Response Theory. *Journal of Personality Assessment*, *94*(4), 427-437. https://doi.org/10.1080/00223891.2012.662188

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research:Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378. https://doi.org/10.1177/1948550617693063

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3).

Garrido, L. E., Barrada, J. R., Aguasvivas, J. A., Martínez-Molina, A., Arias, V. B., Golino, H. F., . . . Rojo-Moreno, L. (2018). Is Small Still Beautiful for the Strengths and Difficulties Questionnaire? Novel Findings Using Exploratory Structural Equation Modeling. *Assessment*, 1073191118780461. https://doi.org/10.1177/1073191118780461

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to Use Broader Internalising and Externalising Subscales Instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British Parents, Teachers and Children. *Journal of Abnormal Child Psychology*, *38*(8), 1179-1191. https://doi.org/10.1007/s10802-010-9434-x

Goodman, R. (2001). Psychometric Properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337-1345. https://doi.org/https://doi.org/10.1097/00004583-200111000-00015

Goodman, R., Meltzer, H., & Bailey, V. (1998). The strengths and difficulties questionnaire: A pilot study on the validity of the self-report version. *European Child & Adolescent Psychiatry*, *7*(3), 125-130. https://doi.org/10.1007/s007870050057

Gunning, F. (1952). *The technique of clear writing*. McGraw-Hill.

Hafekost, J., Lawrence, D., Boterhoven de Haan, K., Johnson, S. E., Saw, S., Buckingham, W. J., . . . Zubrick, S. R. (2016). Methodology of Young Minds Matter: The second Australian Child and Adolescent Survey of Mental Health and Wellbeing. *Australian & New Zealand Journal of Psychiatry*, *50*(9), 866-875. https://doi.org/10.1177/0004867415622270

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

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Humphrey, N., & Wigelsworth, M. (2016). Making the case for universal school-based mental health screening. *Emotional and Behavioural Difficulties*, *21*(1), 22-42. https://doi.org/10.1080/13632752.2015.1120051

Irwing, P., & Hughes, D. J. (2018). Test Development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing* (pp. 3-47). https://doi.org/10.1002/9781118489772.ch1

Jackson, R. H., Davis, T. C., Bairnsfather, L. E., George, R. B., Crouch, M. A., & Gault, H. (1991). Patient reading ability: an overlooked problem in health care. *Southern medical journal*, *84*(10), 1172-1175. https://doi.org/10.1097/00007611-199110000-00004

Janan, D., & Wray, D. (2012). *Readability: the limitations of an approach through formulae.* British Educational Research Association Annual Conference, University of Manchester, http://www.leeds.ac.uk/educol/documents/213296.pdf

Jensen, S. A., Fabiano, G. A., Lopez-Williams, A., & Chacko, A. (2006). The reading grade level of common measures in child and adolescent clinical psychology. *Psychological Assessment*, *18*(3), 346-352. https://doi.org/10.1037/1040-3590.18.3.346

Johnston, C., & Gowers, S. (2005). Routine Outcome Measurement: A Survey of UK Child and Adolescent Mental Health Services. *Child and Adolescent Mental Health*, *10*(3), 133-139. https://doi.org/doi:10.1111/j.1475-3588.2005.00357.x

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel.* https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213-236. https://doi.org/10.1002/acp.2350050305

Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., . . . Zimmermann, J. (2018). Progress in achieving quantitative classification of psychopathology. *World Psychiatry*, *17*(3), 282-293. https://doi.org/10.1002/wps.20566

Lai, K., & Green, S. B. (2016). The Problem with Having Two Watches: Assessment of Fit When RMSEA and CFI Disagree. *Multivariate Behavioral Research*, *51*(2-3), 220-239. https://doi.org/10.1080/00273171.2015.1134306

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936-949. https://doi.org/10.3758/s13428-015-0619-7

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales. *Methodology*, *4*(2), 73-79. https://doi.org/10.1027/1614-2241.4.2.73

Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., & Nagengast, B. (2011). Methodological Measurement Fruitfulness of Exploratory Structural Equation Modeling (ESEM): New Approaches to Key Substantive Issues in Motivation and Engagement. *Journal of*

Psychoeducational Assessment, *29*(4), 322-346. https://doi.org/10.1177/0734282911406657 Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*, *49*(6), 1194-1218. https://doi.org/10.1037/a0026913

Marsh, H. W., Vallerand, R. J., Lafrenière, M.-A. K., Parker, P., Morin, A. J. S., Carbonneau, N., . . . Paquet, Y. (2013). Passion: Does one scale fit all? Construct validity of two-factor passion scale and psychometric invariance over different activities and languages. *Psychological assessment*, *25*(3), 796-809. https://doi.org/10.1037/a0032573

Matsunaga, M. (2010). How to factor-analyze your data right: do's, don'ts, and how-to's. *International journal of psychological research*, *3*(1), 97-110. https://doi.org/https://doi.org/10.21500/20112084.854

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568-592. https://doi.org/10.1037/0021-9010.93.3.568

Moilanen, K. L., Shaw, D. S., & Maxwell, K. L. (2010). Developmental cascades: Externalizing, internalizing, and academic competence from middle childhood to early adolescence. *Development and Psychopathology*, *22*(3), 635-653. https://doi.org/10.1017/S0954579410000337

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*(5), 407-422. https://doi.org/https://doi.org/10.1016/j.intell.2013.06.004

Muthén, B., & Muthén, L. (2016). *warning about parameters and clusters*. Retrieved 01/12/2019 from http://www.statmodel.com/discussion/messages/12/20967.html?1463144022

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2017). *Regression and mediation analysis using Mplus*. Muthén & Muthén Los Angeles, CA.

Muthén, L. K., & Muthén, B. O. (1998-2017). Mplus User's Guide. Eighth Edition. In. Los Angeles, CA: Muthén & Muthén.

Ng Fat, L., Scholes, S., Boniface, S., Mindell, J., & Stewart-Brown, S. (2017). Evaluating and establishing national norms for mental wellbeing using the short Warwick–Edinburgh Mental Well-being Scale (SWEMWBS): findings from the Health Survey for England. *Quality of Life Research*, *26*(5), 1129-1144. https://doi.org/10.1007/s11136-016-1454-8

NHS Digital. (2018). *Mental Health of Children and Young People in England, 2017 Summary of key findings*. https://files.digital.nhs.uk/F6/A5706C/MHCYP%202017%20Summary.pdf

Oakland, T., & Lane, H. B. (2004). Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests. *International Journal of Testing*, *4*(3), 239-252. https://doi.org/10.1207/s15327574ijt0403_3

Omrani, A., Wakefield-Scurr, J., Smith, J., & Brown, N. (2018). Survey Development for Adolescents Aged 11–16 Years: A Developmental Science Based Guide. *Adolescent Research Review*. https://doi.org/10.1007/s40894-018-0089-0

Ortuño-Sierra, J., Chocarro, E., Fonseca-Pedrero, E., Riba, S. S. i., & Muñiz, J. (2015). The assessment of emotional and Behavioural problems: Internal structure of The Strengths and Difficulties Questionnaire. *International Journal of Clinical and Health Psychology*, *15*(3), 265-273. https://doi.org/https://doi.org/10.1016/j.ijchp.2015.05.005

Ortuño-Sierra, J., Fonseca-Pedrero, E., Aritio-Solana, R., Velasco, A. M., de Luis, E. C., Schumann, G., . . . consortium, I. (2015). New evidence of factor structure and measurement invariance of the SDQ across five European nations. *European Child & Adolescent Psychiatry*, *24*(12), 1523-1534. https://doi.org/10.1007/s00787-015-0729-x

Panayiotou, M., Humphrey, N., & Wigelsworth, M. (2019). An empirical basis for linking social and emotional learning to academic performance. *Contemporary Educational Psychology*, *56*,193-204. https://doi.org/https://doi.org/10.1016/j.cedpsych.2019.01.009

Patalay, P., Hayes, D., & Wolpert, M. (2018). Assessing the readability of the self-reported Strengths and Difficulties Questionnaire. *BJPsych Open*, *4*(2), 55-57. https://doi.org/10.1192/bjo.2017.13

Percy, A., McCrystal, P., & Higgins, K. (2008). Confirmatory factor analysis of the adolescent self-report Strengths and Difficulties Questionnaire. *European Journal of Psychological Assessment, 24*(1), 43–48. https://doi.org/10.1027/1015-5759.24.1.43

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

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Raykov, T., & Marcoulides, G. A. (2016). Scale Reliability Evaluation Under Multiple Assumption Violations. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 302-313. https://doi.org/10.1080/10705511.2014.938597

Saris, W. E. (2014). *Design, evaluation, and analysis of questionnaires for survey research* (Second edition. ed.). Wiley.

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating Model Fit With Ordered Categorical Data Within a Measurement Invariance Framework: A Comparison of Estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167-180. https://doi.org/10.1080/10705511.2014.882658

Suárez-Alvarez, J., Pedrosa, I., Lozano Fernández, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema*, *30*(2), 149-158. https://doi.org/10.7334/psicothema2018.33

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., . . . de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34-42. https://doi.org/https://doi.org/10.1016/j.jclinepi.2006.03.012

Tóth-Király, I., Bõthe, B., Rigó, A., & Orosz, G. (2017). An Illustration of the Exploratory Structural Equation Modeling (ESEM) Framework on the Passion Scale. *8*(1968). https://doi.org/10.3389/fpsyg.2017.01968 van de Looij-Jansen, P. M., Goedhart, A. W., de Wilde, E. J., & Treffers, P. D. A. (2011). Confirmatory factor analysis and factorial invariance analysis of the adolescent self-report Strengths and Difficulties Questionnaire: How important are method effects and minor factors? *British Journal of Clinical Psychology*, *50*(2), 127-144. https://doi.org/10.1348/014466510x498174

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain. *PLOS ONE*, *8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

Vostanis, P. (2006). Strengths and Difficulties Questionnaire: Research and clinical applications. *Current Opinion in Psychiatry*, *19*(4), 367-372. https://doi.org/10.1097/01.yco.0000228755.72366.05

Weijters, B., & Baumgartner, H. (2012). Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research*, *49*(5), 737-747. https://doi.org/10.1509/jmr.11.0368

Wigelsworth, M., Humphrey, N., & Lendrum, A. (2012). A national evaluation of the impact of the secondary social and emotional aspects of learning (SEAL) programme. *Educational Psychology*, *32*(2), 213-238. https://doi.org/10.1080/01443410.2011.640308

Wolpert, M., Cheng, H., & Deighton, J. (2015). Measurement Issues: Review of four patient reported outcome measures: SDQ, RCADS, C/ORS and GBO – their strengths and limitations for clinical use and service evaluation. *20*(1), 63-70. https://doi.org/doi:10.1111/camh.12065

Implications of Paper 3 for Policy Makers and/or Educators and Teachers

While Papers 1 and 2 suggested general caution was necessary when considering and comparing positive and negative mental health outcomes, Paper 3 has a very specific implication: The self-report SDQ should not be used or recommended. The structural issues mean it cannot reasonably be scored, and the item quality and readability findings further call into question its validity. The paper makes clear the consequences of omitting proper scale development practices and suggests investment here should be a priority. Instead measures/subscales, such as those highlighted as having at least content and structural validity evidence in Paper 4 (e.g., KIDSCREEN) should be considered.

Paper 4: Measuring General Mental Health in Early-Mid Adolescence: A Systematic Meta-Review of Content and Psychometrics

Status: conditionally accepted, submitted to JCPP Advances December 2021

Black, L., Panayiotou, M., & Humphrey, N. (under review). Measuring general mental health in early-mid adolescence: A systematic meta-review of content and psychometrics. *JCPP Advances*. doi:10.31234/osf.io/e3y8r

Supplementary material can be found at https://osf.io/k7qth/

Measuring General Mental Health in Early-Mid Adolescence: A Systematic Meta-Review of Content and Psychometrics

Louise Black, Margarita Panayiotou, Neil Humphrey

The University of Manchester

Declaration of Interest: The authors declare no competing interests.

PROSPERO registration: CRD42020184350

https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020184350

Abstract

Background

Adolescent mental health is a major concern and brief general self-report measures can facilitate insight into intervention response and epidemiology via large samples. However, measures' relative content and psychometrics are unclear.

Method

A systematic search of systematic reviews was conducted to identify relevant measures. We searched PsycINFO, MEDLINE, EMBASE, COSMIN, Web of Science, and Google Scholar. Theoretical domains were described, and item content was coded and analysed, including via the Jaccard index to determine measure similarity. Psychometric properties were extracted and rated using the COSMIN system.

Results

We identified 22 measures from 19 reviews, which considered general mental health (positive and negative aspects together), life satisfaction, quality of life (mental health subscales only), symptoms, and wellbeing. Measures were often classified inconsistently within domains at the review level. Only 25 unique indicators were found and several indicators were found across the majority of measures and domains. Most measure pairs had low Jaccard indexes, but 6.06% of measure pairs had >50% similarity (most across two domains). Measures consistently tapped mostly emotional content but tended to show thematic heterogeneity (included more than one of emotional, cognitive, behavioural, physical and social items). Psychometric quality was generally low.

Conclusions

Brief adolescent general mental health measures have not been developed to sophisticated standards, likely limiting robust inferences. Researchers and practitioners should attend carefully to specific items included, particularly when deploying multiple measures. Key considerations, more promising measures, and future directions are highlighted.

Keywords: adolescence, measurement, mental health

Introduction

Adolescence, the phase starting around age 10 (Sawyer et al., 2018), appears pivotal for mental health problems, playing host to the first onset of the majority of lifetime cases (Jones, 2013). There is also evidence mental health of young people is worse than in previous generations (Collishaw, 2015). Despite a striking need to improve our understanding of mental health in this age group, research has typically faced major methodological problems, including low statistical power, poor measurement, and analytical flexibility (Rutter & Pickles, 2016). High-quality research going forward will likely be underpinned by well-developed brief measures to facilitate large samples. This meta-review focuses on the content and psychometric properties of self-report measures to aid researchers and practitioners in selecting indicators and measures more likely to lead to valid inferences. Various operationalizations of general mental health (GMH) exist (e.g., disorders or wellbeing). However, it is currently unclear how these constructs relate to one another conceptually or their relative psychometric qualities. Reviews have been conducted considering general measures including multiple operationalizations (Bentley et al., 2019; Deighton et al., 2014). These inevitably have different criteria, definitions, resulting measures, and ratings of psychometric properties. It is crucial to bring this work together to make clear which brief measures are considered to measure GMH. Consistent and robust assessment of psychometric results can then also be applied.

Existing reviews have also not assessed item content (e.g. the symptoms, thoughts, behaviours and experiences that are considered by measures). This is a key omission. For instance, some researchers and practitioners may have clear theories about why one domain of GMH in particular is of interest (e.g., affected by an intervention). However, without explicit attention to content, results may be selected in a more data driven way. While it is the norm to register primary outcomes in trials, in adolescent mental health, some recommend multiple measures are explored for sensitivity (Horowitz & Garber, 2006). Observational studies also often collect multiple similar domains (e.g., NHS Digital, 2018). While such exploratory approaches play an important role and flexibility can occur even after registration (Scheel et al., 2020), we suggest the content of measures should be attended to, particularly when combined. Before inferences are made about constructs, we must gain better understanding of how measures relate conceptually.

This is also vital given the noisiness of adolescent mental health data (Wolpert & Rutter, 2018). Consider a case where a symptom measure (e.g. depression) shows significant improvement after intervention but a wellbeing measure does not. If the wellbeing measure covers theoretically distinct content this is likely to be a robust finding. However, if both cover depression, affect or other indicators which could appear in either domain (Alexandrova & Haybron, 2016), this is less likely to be the case.

While analysis of item content is lacking, there is literature describing the theoretical domains to which measures belong. For instance, measures may be based on diagnostic systems such as the Diagnostic and Statistical Manual of Mental Disorders or frameworks such as hedonic or eudaimonic wellbeing (Ryan & Deci, 2001). However, we chose to focus on item rather than construct mapping for several reasons: First, it is a known problem that measures with different labels sometimes measure the same construct, while others with the same label measure different constructs (jingle-jangle fallacy; Marsh, 1994). Second, measures and their sub-domains are often heterogeneous (Newson et al., 2020). Third, psychometric validations can be data-driven, resulting in items with beneficial statistical properties prioritized over those considered to be theoretically key (Alexandrova & Haybron, 2016; Clifton, 2020). We therefore argue against further reification of construct boundaries.

To aid comparison there have also been calls for common measures (Wolpert, 2020). However, a key problem is that different measures are likely appropriate for different contexts (Patalay & Fried, 2020). We argue the choice of measures for individual studies, or to standardize across studies, should be informed by analyses such as those reported here.

Method

A systematic search was conducted to identify measures following PRISMA guidelines (see supporting information). We registered a number of research questions which considered: which theoretical domains were included in GMH (RQ1); the number of unique indicators (RQ2); the presence of key common indicators across measures/domains (RQ3); the proportions of items assessing broader themes (cognitive/affective/behavioural/physical) by measure/domain (RQ4); which measures best represent common indicators (RQ5); the similarity of measures within and between domains (RQ6); measures' time frames (RQ7); psychometric properties (RQ8); statistical and conceptual consistency (RQ9).

To answer these we defined several units of analysis. First, we use the term *theoretical* *domains* to refer to constructs described at the review level (e.g. life satisfaction). We grouped included reviews into theoretical domains inductively. Second, we use *indicator* to refer to specific question types capturing individual symptoms, thoughts, behaviours or experiences (e.g. sadness). Finally, we use *broad themes* to classify whether items tapped emotional, physical, social, cognitive or behavioural content.

Full search terms, eligibility criteria, inter-rater reliability information, indicator codes, and R scripts are provided on the Open Science Framework (https://osf.io/k7qth/) and in the supporting information. The COSMIN database of systematic reviews of measures was searched, as well as PsycINFO, MEDLINE, EMBASE, Web of Science, and Google Scholar. Reference lists of eligible studies were also searched. Search terms relating to the population (e.g., adolescen* OR youth*, etc.), measurement (e.g., survey* OR questionnaire*, etc.), and construct of interest (e.g., "mental health" OR wellbeing, etc.) were combined using the AND operator. Where databases allowed, hits were limited to reviews, and English, since we aimed to review English-language measures validated with English speakers.

To appraise the methodological quality of reviews from which we drew measures, we employed the quality assessment of systematic reviews of outcome measurement instruments tool (see Supplementary Table 1; Terwee et al., 2016).

A subset of measures were initially discussed by all authors as the basis for the coding strategy. We aimed to code at a semantic level. However, given we could not be blind to the intended content of measures (e.g., measures' titles could give this away), coding could not be entirely inductive (Braun & Clarke, 2006). A hybrid approach allowed initial coding to be either specific or broad, with some codes collapsed into more general categories in subsequent coding, and others split up. After the initial meeting, the first author generated a full set of preliminary codes for all included items which were reviewed by the other authors. These were refined into a final set through discussion. In the final coding (https://osf.io/k7qth/), we aimed to collapse as much as possible without losing information. This was to avoid false positive differences between measures (Newson et al., 2020). Wherever possible, items were given a single code, but for items assessing more than one experience (e.g. sadness and worry), two codes were assigned. Each item was also assigned one or more broad themes (e.g., losing sleep over worry was considered physical and emotional).

As has been used elsewhere, similarity between measures was calculated via the Jaccard index (Fried, 2017). This index is the number of common indicators divided by the total number of indicators across a pair of measures. Each measure therefore gains a 1 or 0 for presence or absence of the indicator (regardless of frequency), making the index unweighted. This was desirable to avoid biased construct dissimilarity through our strategy of including whole measures for domains like symptoms, but shorter subscales from quality of life. Items with double codes were both included as indicators for a given measure.

Though we initially intended to conduct secondary searches for psychometric evidence (Black, Panayiotou, et al., 2020), we instead opted to use primary studies cited in reviews. This was more feasible, was supported by the quality of reviews (see Supplementary Table 1), and frequent inclusion of measures in several reviews (see Figure 2). We reported only psychometric properties analysed in samples consistent with our criteria (e.g., not clinical samples or other age ranges) and included only studies reporting on relevant COSMIN elements at the level we considered (subscales or whole measures). All references and raw psychometric information extracted can be found at https://osf.io/k7qth/.

We used the COSMIN rating system for psychometric properties (Mokkink et al., 2018), which recommends consideration of content validity, structural validity, internal consistency, measurement invariance, reliability, measurement error, hypothesis testing for construct validity, responsiveness, and criterion validity. A few adaptations were necessary in the current study and are described in the supporting information. The rating takes the form: +, -, +/- (inconsistent), ? (indeterminate), and where no information was available we rated no evidence (NE).

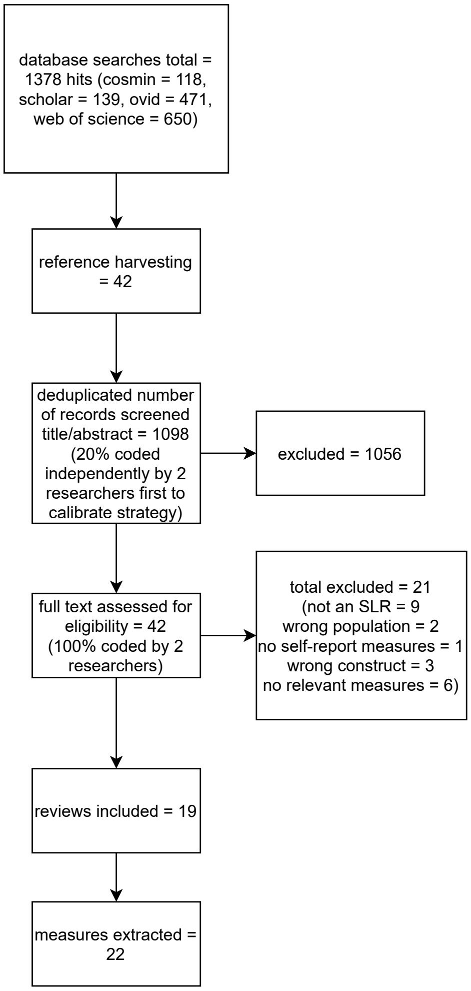
In order to address RQ9, we assessed whether measures/subscales were conceptually homogenous (H). We considered homogeneity to be present where only one broad theme was assessed. This was combined with statistical consistency (S), which we considered to be present where measures scored at least +/- for both structural validity and internal consistency. Measures could therefore be H+S+, H-S+, H-S+, or H-S-.

Results

A flowchart of the review stages is presented in Figure 1 with the primary reason for exclusion reported for full-texts. The number of measures corresponds to collapsing different versions of the same measure, and subscales within measures are not counted separately.

Figure 1.

Flow Diagram of Review Process.


database searches total = 1378 hits (cosmin = 118, scholar = 139, ovid = 471, web of science = 650)
reference harvesting = 42
deduplicated number of records screened title/abstract = 1098 (20% coded independently by 2 researchers first to calibrate strategy)
excluded = 1056
full text assessed for eligibility = 42 (100% coded by 2 researchers)
total excluded = 21 (not an SLR = 9 wrong population = 2 no self-report measures = 1 wrong construct = 3 no relevant measures = 6)
reviews included = 19
measures extracted = 22

Results of the quality assessment indicate mixed quality (see Supplementary Table 1). For instance, the vast majority of studies (94.74%) defined the construct of interest, and used multiple databases. However, reviewing, quality assessment and extraction of psychometric properties were often not clearly reported or were conducted only by a single researcher. Results are therefore in line with the general field of measure reviews (Terwee et al., 2016).

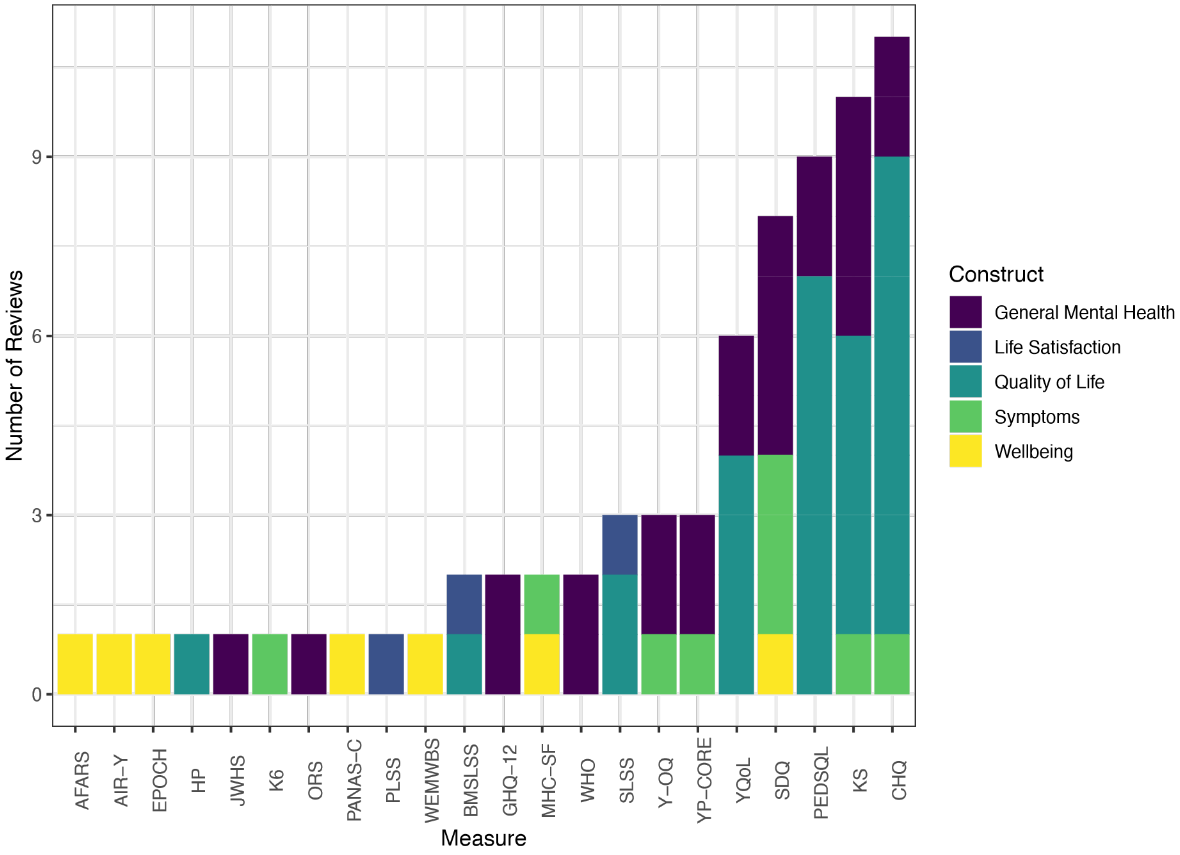
We included all criteria set out by Terwee et al. (2016). While 100% of studies reported the population of interest, since we had a specific age criterion, several reviews explicitly noted developmental considerations (Harding, 2001; Janssens, Thompson Coon, et al., 2015; Kwan & Rickwood, 2015; Rose et al., 2017), suggesting this had been appropriately considered.

The 19 reviews covered five theoretical domains (RQ1): general mental health, holistic approaches including positive and social aspects (Bentley et al., 2019; Bradford & Rickwood, 2012; Kwan & Rickwood, 2015; Wolpert et al., 2008); symptoms (Becker-Haimes et al., 2020; Deighton et al., 2014; Stevanovic et al., 2017); quality of life, including functional disability and patient reported outcome measures (Davis et al., 2006; Fayed et al., 2012; Harding, 2001; Janssens, Rogers, et al., 2015; Janssens, Thompson Coon, et al., 2015; Rajmil et al., 2004; Ravens-Sieberer et al., 2006; Schmidt et al., 2002; Solans et al., 2008; Upton et al., 2008); wellbeing, positively-framed strengths-based measures or those with substantial proportions of positive items/subscales (Rose et al., 2017; Tsang et al., 2012); and life satisfaction (Proctor et al., 2009). Figure 2 demonstrates some measures appeared in several reviews under different domains (e.g., Child Health Questionnaire, CHQ and KIDSCREEN). Given the lack of consensus among reviews about which constructs measures fell under, we regrouped measures based on descriptions in validation papers cited in reviews. This resulted in the following which were used to inform subsequent research questions: affect, life satisfaction (LS), quality of life (QoL), symptoms, and wellbeing.[5]

[5] Though we treated these as a single group, QoL measures were noted to include subscales for the following domains that met our criteria: symptoms (CHQ, KIDSCREEN and PedsQL), wellbeing (KIDSCREEN), life satisfaction (Healthy Pathways, HP and Youth Quality of Life, YQoL), and psychological QoL (KIDSCREEN and WH-QoL), which contained a mixture of positive and negative indicators.

Figure 2

Summary of Measures and Reviews



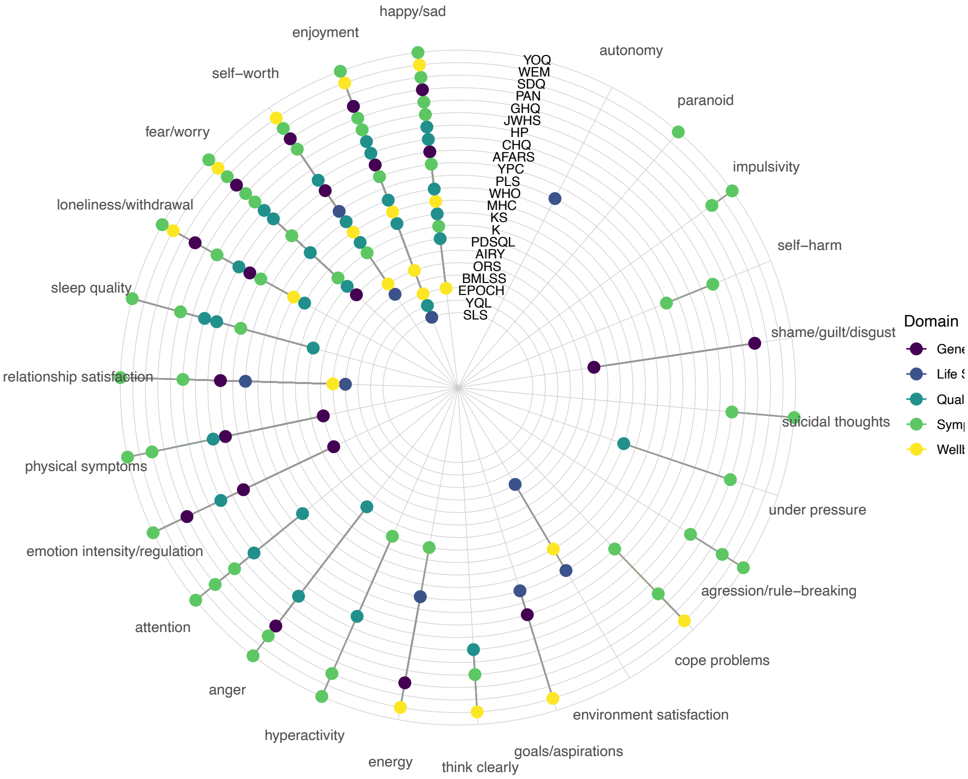
Note. Measures' full names can be found in Table 2.

The measures extracted from reviews are presented in supplementary table 2, including measure time-frames (RQ7). For the 14 measures with clear time frames, all but one considered periods of one to four weeks.

Our initial coding generated 45 codes which were collapsed into a final set of 25 (RQ2, see Figure 3 and https://osf.io/k7qth/). For example, emotion intensity/regulation covered getting upset easily/impatience/strong positive and negative emotional responses/excited. Since we had 285 items, the reduction to indicators was 91.23%.

Figure 3.

25 Indicators Across Measures by Domain



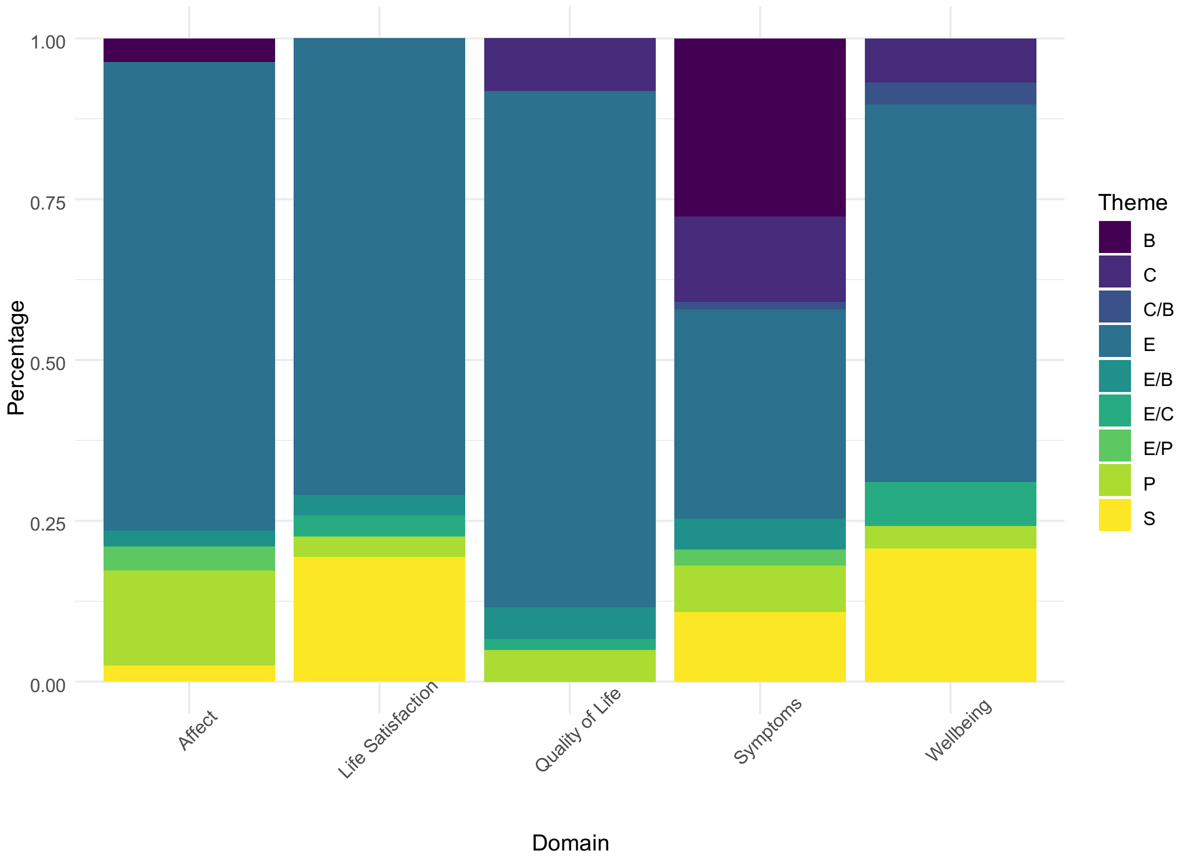
Note. Measures’ full names can be found in Table 2.

Five items had two codes applied, resulting in three additional indicators being allocated to measures. The indicators in Figure 3 are ordered by how commonly they occur across measures, with happy/sad and enjoyment both occurring in 72.72% of measures, and autonomy and paranoid occurring across 4.54% (RQ3). The outer-most measure, YOQ, has the most indicators while SLS in the centre of the plot has the least. Symptom measures covered the most indicators (84%), and LS measures the least (28%). The other domains each covered roughly half of all indicators.

Broader-level themes are shown in Figure 4 (RQ4). These were not hierarchical but coded per item. Items within the same indicator often but not always had the same broad theme, reflecting our to collapse initial indicator codes as much as possible. For instance, 11 of the loneliness/withdrawal items were coded as tapping social content (e.g., “I withdraw from my family and friends”) while the remaining 5 were emotional (e.g., “feel lonely”). The majority of indicators tapped emotional experiences. Symptom measures had a higher proportion of behavioural and cognitive indicators, reflecting more coverage of externalizing problems.

Figure 4

Broader Themes Across Domains



Note.

B = behavioural

C= cognitive

E = emotional

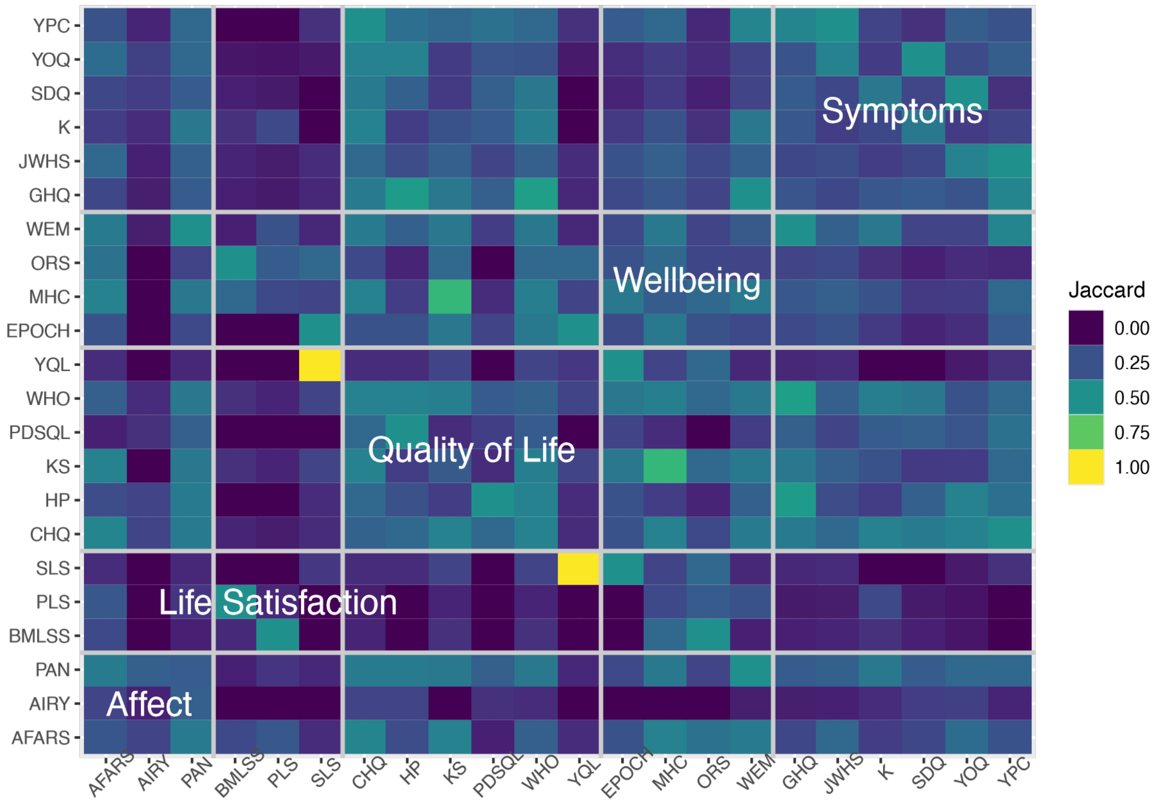
P = physical

S = social

Overlap between pairs of measures ranged from 0-1 (*M* = .23, *SD* = .15). Only 14 (6.06%) of measure pairs had similarity >.50 (Figure 5, RQ5). Of these, 10 (4.33% of all pairs) were for pairs of measures from different domains. LS measures typically had low overlap with other measures. Affect measures also seemed to have relatively lower overlap while the remaining domains showed similar overlap. Average similarity for each measure with all others (shown on the diagonal of Figure 5) ranged from .09 (AIR-Y) to .32 (CHQ), *M* = .23, *SD* = .06. Measures with higher average overlap were typically wellbeing and QoL instruments. The pair of measures with perfect overlap (SLS and YQL), cover only enjoyment.

Figure 5

Jaccard Index by Measure



Note. Measures' full names can be found in Table 2.

Measures appeared slightly more similar within than between domains. This can be seen by comparing the large diagonal boxes marked with domains in Figure 5 to other pairs in each row/column marked by pale grid lines (see also Table 1 for averaged Jaccard Index by domain). We found no more than 42% similarity between measures of the same domain. Symptoms and LS were particularly dissociated.

Table 1

Average Jaccard Indexes Within (Diagonal) and Between (Lower Triangular) Domains

	Affect	Life Satisfaction	Quality of Life	Wellbeing	Symptoms
Affect	0.33				
Life Satisfaction	0.11	0.33			
Quality of Life	0.24	0.13	0.42		
Wellbeing	0.24	0.24	0.30	0.38	
Symptoms	0.24	0.08	0.30	0.23	0.42

The psychometric properties of measures are shown in Table 2 (RQ8). There was no evidence available for measurement error for any measure so this was omitted. Six measures (27.27%) scored positively for content validity, a fundamental property (Mokkink et al., 2018). These measures all also scored favourably for construct validity, though no further positive results were found for these, suggesting overall low quality. HS scores are shown in Table 2.

Discussion

This study systematically brought together measures across domains identified in systematic reviews as capturing adolescent GMH and is the first, to our knowledge, to consider content and psychometrics together. The current paper affords several new insights: First, theoretical domains were inconsistent, with individual measures frequently considered to belong several. Second, despite a relatively large number of measures and domains, we found these to be captured by only 25 indicators, with some appearing across the majority of measures/domains. Third, this narrow range was echoed in broader themes with most featuring emotional content. Fourth, quantitative analysis of measure overlap suggested only a few pairs of measures were highly similar, but these were largely for pairs from different domains. Finally, though we considered measures/subscales that were recommended for sum scoring, we found only a few with theme-level homogeneity, and fewer still which also showed statistical coherence. These findings suggest brief measurement of adolescent GMH is relatively unsophisticated. Researchers and practitioners should therefore be cautious when selecting, analysing, and interpreting such measures, particularly if considering multiple outcomes. In the following sections we highlight particular considerations.

Table 2

COSMIN Ratings of Measures and HS Scores

Measure Domain	Measure (full name)	Content Validity	Structural Validity	Internal Consistency	Reliability	Construct Validity	Measurement Invariance	Broad Themes	HS Score
Symptoms	GHQ-12 (General Health Questionnaire)	-	+	+	NE	+	NE	4	H-S+
	SDQ (Strengths and Difficulties Questionnaire)	NE	+/-	-	?	-	-	Conduct = 2 Emotional = 3 Hyperactivity = 2 Total = 4	H-S-
	YP-CORE (Young Person Clinical Outcomes in Routine Evaluation)	+	NE	?	?	+	NE	5	H-S-

	YOQ (Youth Outcome Questionnaire)	+	NE	?	?	+	NE	5	H-S-
	K (6) (Kessler)	-	NE	?	NE	-	NE	3	H-S-
	JWHS-76 (Juvenile Wellness and Health Survey)	+	NE	?	NE	+	NE	4	H-S-
Quality of Life	KS (KIDSCREEN)	+	+/-	?	-	+	+	Moods and emotions= 2 Psychological wellbeing = 2	H-S-

PedsQL (Pediatric Quality of Life Inventory)	?	NE	?	NE	+	NE	2	H-S-
CHQ (Child Health Questionnaire)	?	NE	?	NE	+	NE	3	H-S-
YQoL-R (Youth Quality of Life Research version)	?	NE	?	+	+	NE	1	H+S-
WHOQOL-BREF	-	NE	?	NE	+	NE	2	H-S-

	(World Health Organization Quality of Life, brief)								
	HP (Healthy Pathways)	+	-	?	NE	+	+	life satisfaction = 1 emotional comfort = 1 negative stress reaction = 2	H+S- / H-S-
	ORS (Outcome Rating Scale)	NE	NE	?	-	+	NE	2	H-S-
Wellbeing									
	EPOCH (Engagement, Perseverance,	NE	+/-	+	?	-	+	1	H+S+

	Optimism, Connectedness, Happiness)								
	WEMWBS (Warwick-Edinburgh Mental Wellbeing Scale)	-	+*	+	-	+	NE	5	H-S+*
	MHC-SF (Mental Health Continuum Short Form)	NE	-	?	NE	+	NE	emotional wellbeing = 1, psychological wellbeing = 3	H+S- / H-S-
Affect	AFARS (Affect and Arousal Scale)	NE	+	+	?	-	NE	Negative affect = 1 Positive affect = 3	H+S+ / H-S+

AIR-Y (Affect Intensity and Reactivity Measure, Youth)	-	-	?	?	?	-	Physiological = 1 Positive affect = 3 Negative reactivity = 2 Negative intensity = 1	H-S- / H+S-
PANAS-C (Positive and Negative Affect Scale, Child)	+	NE	?	NE	+	NE	Positive affect = 3 Negative affect = 2	H-S-
PLSS	NE	NE	?	NE	+	NE	4	H-S-

Life Satisfaction	(Perceived Life Satisfaction Scale)								
	SLSS (Student Life Satisfaction Scale)	NE	NE	?	?	+	NE	2	H-S-
	BMSLSS	NE	NE	?	?	+	NE	2	H-S-

	(Brief Multidimensional Student Life Satisfaction Scale)

* *Note*. Many residual correlations were added, likely driving up fit.

HS = conceptual Homogeneity Statistical consistency score.

A few indicators stood out as appearing in >50% of measures and 80-100% domains: happy/sad, enjoyment, fear/worry, and self-worth. This suggests these may be broadly useful, since validation processes have frequently led to their inclusion as indicators of GMH. Common indicators may also explain the classification inconsistency of measures into domains. While symptom measures had more idiosyncratic indicators, likely reflecting indicators that could only be framed negatively (e.g., suicidal thoughts), LS had the narrowest range of indicators. Despite this, some LS measures had relatively high thematic heterogeneity (see Table 2), likely reflecting that LS considers satisfaction across a range of areas (e.g. social and emotional). However, our findings suggest this breadth should not be considered indicative of GMH. The validity of findings (particularly external) aiming to capture GMH via LS may therefore be threatened.

The percentage reduction from items to indicators seen here, 91.23%, was greater than in similar studies (of single disorders), where the number of items was typically reduced by 45.3-77.3% (Chrobak et al., 2018; Fried, 2017; Hendriks et al., 2020; Visontay et al., 2019). The percentage reduction inevitably reflects how conservative coding was, though all studies described being cautious. We also saw the full range of overlap at the measure level whereas the aforementioned studies had smaller ranges (.26-.61). We found some pairs of measures (4.33%), from domains labelled as being different, had >50% overlap in terms of content, suggestive of the jangle fallacy. Generally though, similarity was low, even within domains, suggesting domains are poorly defined. Researchers and practitioners should therefore attend to the specific items in questionnaires before deploying them, drawing on experts and analyses such as that presented here.

Not doing so could create problems with analysis and interpretation. For instance, even though we targeted subscales and measures designed to directly assess mental states, rather than antecedents, indicators of wider functioning were nevertheless included, such as relationships and aspirations. If GMH measures include such indicators, then careful treatment is needed when analysing potentially overlapping correlates. Relatedly, while some have called for measures of functioning (e.g., QoL) to be consistently used to help compare studies (Mullarkey & Schleider, 2021), our analysis suggests that what constitutes mental health-related QoL or functioning is not consistently defined.

In terms of standardizing measurement for capturing the range of GMH, no single measure or domain represented the entire spectrum. As discussed, we aimed to collapse codes wherever possible, emphasizing the starkness of this finding. The measures with the highest number of broad themes (see Table 2), also tended to have the most indicators (e.g., YOQ had the most with 15 while GHQ, WEMWBS, PANAS and SDQ all had nine, see Figure 3). However, these measures did not share the same indicators, with the greatest similarity between YOQ and SDQ at 50% (see Figure 5, code and data, https://osf.io/k7qth/). The inconsistency found at the review level is therefore reflected in our content findings: In terms of content, measures within theoretical domains are mostly not interchangeable, while some typically understood to capture different domains could be. This is of vital significance given the leap usually made from measure to construct when discussing findings, and makes clear potential problems of generalizability (Yarkoni, 2020).

Psychometric evidence was frequently lacking and COSMIN scores were low. Our results also confirm the general tendency to report only basic structural evidence (Flake et al., 2017). Though construct validity was frequently reported and positive, it should be treated with some caution since it has been suggested the type considered in the COSMIN rubric may not be valid if content and structural validity have not been considered (Flake et al., 2017), as was often the case here. Of the measures which scored positively for content validity, only KIDSCREEN and EPOCH evaluated structural validity, scoring +/- and – respectively. LS seemed particularly psychometrically problematic. QoL and outcome-focused symptom measures showed better content validity.

As noted above, statistical coherence was typically unclear or poor. Though measures/subscales were recommended for sum scoring, they tended to cover more than one broad theme, suggesting conceptual unidimensionality was untenable. It is likely measures/constructs with thematic heterogeneity are not well suited to internal consistency metrics or sum scoring (Fried & Nesse, 2015). Similarly, reliability should only be prioritised by developers within theoretical units since otherwise statistical reliability can be introduced via wording or other artefacts, rather than structural validity (Clifton, 2020).

Most measures scored H-S-. We recommend such measures are not sum scored since this is not supported theoretically or statistically. Heterogeneous constructs may be desirable, particularly for GMH given one of its highlighted benefits is to provide broad insight (Deighton et al., 2014). We therefore question the logic of sum scores in this area. While items from measures included in this review could therefore provide insight into GMH via methods other than sum scoring (e.g., network models), further work is needed to support such approaches.

GHQ-12, WEMWBS and AFARS positive affect all scored H-S+. This could be interpreted in several ways. It is possible these measures represent constructs that can be assessed from a variety of perspectives (indeed, positive affect was consistently heterogeneous). H-S+ could also signal data-driven development without adequate consideration of whether sum scoring is theoretically appropriate. S+ could be the result of post-hoc model modifications: In the case of WEMWBS, the addition of 28 error correlations in the adolescent validation is a potential cause for concern since it is unlikely these would be added if not needed to drive up model fit.[6] Similarly, none of these three measures met our threshold for content validity with GHQ-12 and WEMWBS both scoring poorly (-) as they were developed for adults. These considerations demonstrate the value of considering theoretical criteria alongside statistical properties. Our novel consideration of conceptual/statistical coherence offers a basis for doing so.

Various subscales (YQoL-R, HP, AIR-Y) scored H+S-. Unless other measures cannot provide adequate indicators, we suggest these should be treated with caution since they could have interpretability or other problems. For instance, age appropriateness can be a particular concern and may drive down psychometric properties (Black, Mansfield, et al., 2020).

Only EPOCH (happiness subscale) and AFARS (negative affect) scored H+S+. These subscales are likely more appropriate for sum scoring. However, the cost of this benefit is fewer GMH indicators (EPOCH contains four, and AFARS negative affect three). Additionally, these measures are by no means likely ideal in all scenarios. In particular, they are both potentially limited by not scoring positively for content validity. Our HS scoring system should therefore not be used to rank measures but be considered alongside issues such as indicators of interest and analytical approach.

This study systematically drew on a large body of systematic reviews, and therefore provides broad coverage of relevant measures and their properties. While some work has provided robust psychometric evaluation (Bentley et al., 2019), this was at the study level, while we were able to combine studies to provide more comprehensive ratings. We also went beyond previous work by

[6] The primary validation includes 28 parameters for residual correlations and does not report fit before the inclusion of these (see https://osf.io/k7qth/).

considering in detail which elements of QoL were relevant to GMH, rather than providing information at the measure level (i.e. general QoL) as has been done previously (e.g., Deighton et al., 2014). We therefore provide novel insight into the specific conceptual overlap of QoL subdomains with other domains of GMH, as well as which subscales can be extracted and scored.

The current study provides a wealth of information for researchers and practitioners. Given the scope of such a project, some compromises were made. First, we were unable to conduct secondary searches for validation studies and therefore relied on the quality of searches conducted in reviews. Since we did not conduct secondary searches ourselves, we cannot be certain relevant papers were not missed. However, our meta-review strategy meant that measures were picked up in multiple reviews (see Figure 2). Second, we did not assess potential methodological bias in validation papers, but rather rated only psychometric quality, for feasibility. Third, our assessment of homogeneity was somewhat crude. However, we based this on broader themes rather than indicators to take into account relationships between indicators. Considering themes rather than indicators was therefore conservative and less likely to underestimate homogeneity and appropriateness for sum scoring.

Conclusion and Recommendations

Though we found a range of constructs defined within GMH and reviews did not always agree which of these individual measures covered, we found a relatively small set of indicators. This relative homogeneity, compared to e.g., depression measures (Fried, 2017), was also seen in measurement time frames and that most items considered emotional content, whereas work looking at disorder measures found greater heterogeneity for these aspects (Newson et al., 2020). This suggests GMH could be assessed briefly. Despite this, while measures within domains showed slightly higher average similarity than pairs across domains, similarity between measures tended to be low, and no measure or domain represented the entire spectrum of indicators.

Findings suggest GMH is not well defined and well-developed measures are lacking. We therefore recommend that where assessment of GMH is the goal, new measures be developed, or existing ones revised. Our review provides excellent groundwork for this by identifying the range of indicators that are likely theoretically relevant. Such analysis has been used to develop general measures for adults (Newson & Thiagarajan, 2020). Our additional assessment of psychometric information, would allow future work to ‘open up’ the codes found in measures with better content validity. This would allow consideration of item types within indicators developed in consultation with stakeholders. For instance, our happy/sad code appeared in most measures but the particular operationalization of this going forward should preference measures which showed some content validity evidence.

In terms of selecting domains, symptom measures captured a broader range likely because some symptoms do not have theoretical positive poles. Researchers and practitioners should therefore consider whether theoretical breadth is important, whether the individual items are of interest, and whether they wish to sum score (this is problematic for diverse item sets). Our findings also underscore that a single measure cannot be selected to represent any domain (given inconsistency within these). However, in terms of psychometrics, the following measures had at least evidence of content and construct validity: YP-CORE, JWHS-76 and YOQ (symptoms), KIDSCREEN (QoL), and PANAS-C (affect). It is difficult to determine the relative psychometric quality of wellbeing measures reviewed given the lack of content validity evidence, though EPOCH (happiness) may be promising, given its match between conceptual and statistical coherence. From a GMH perspective, we recommend LS measures are avoided as these are psychometrically the weakest and show poorer coverage of GMH indicators. We recommend researchers and practitioners considering measures we reviewed draw on our code and data to assess specific content and properties relative to their context. Finally, our analysis suggests that researchers should not combine measures from different domains without accounting for likely covariance, and acknowledging potential systematic overlap due to common content.

References

Alexandrova, A., & Haybron, D. M. (2016). Is Construct Validation Valid? *Philosophy of Science*, *83*(5), 1098-1109. https://doi.org/10.1086/687941

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

Bentley, N., Hartley, S., & Bucci, S. (2019). Systematic Review of Self-Report Measures of General Mental Health and Wellbeing in Adolescent Mental Health. *Clinical Child and Family Psychology Review*, *22*(2), 225-252. https://doi.org/10.1007/s10567-018-00273-x

Black, L., Mansfield, R., & Panayiotou, M. (2020). Age Appropriateness of the Self-Report Strengths and Difficulties Questionnaire. *Assessment*, *0*(0), 1073191120903382. https://doi.org/10.1177/1073191120903382

Black, L., Panayiotou, M., & Humphrey, N. (2020). *Item-level analysis of recommended self-report measures: what are the indicators of adolescent general mental health?* PROSPERO 2020. https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020184350

Bradford, S., & Rickwood, D. (2012). Psychosocial assessments for young people: a systematic review examining acceptability, disclosure and engagement, and predictive utility. *Adolescent health, medicine and therapeutics*, *3*, 111-125. https://doi.org/10.2147/AHMT.S38442

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Chrobak, A. A., Siwek, M., Dudek, D., & Rybakowski, J. K. (2018). Content overlap analysis of 64 (hypo)mania symptoms among seven common rating scales. *International Journal of Methods in Psychiatric Research*, *27*(3), e1737. https://doi.org/10.1002/mpr.1737

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, *25*(3), 259-270. https://doi.org/10.1037/met0000236

Collishaw, S. (2015). Annual Research Review: Secular trends in child and adolescent mental health. *Journal of Child Psychology and Psychiatry*, *56*(3), 370-393. https://doi.org/10.1111/jcpp.12372

Davis, E., Waters, E., Mackinnon, A., Reddihough, D., Graham, H. K., Mehmet-Radji, O., & Boyd, R. (2006). Paediatric quality of life instruments: a review of the impact of the conceptual framework on outcomes. *Developmental Medicine & Child Neurology*, *48*(4), 311-318. https://doi.org/https://doi.org/10.1017/S0012162206000673

Deighton, J., Croudace, T., Fonagy, P., Brown, J., Patalay, P., & Wolpert, M. (2014). Measuring mental health and wellbeing outcomes for children and adolescents to inform practice and policy: a review of child self-report measures. *Child and Adolescent Psychiatry and Mental Health*, *8*(1), 14. https://doi.org/10.1186/1753-2000-8-14

Fayed, N., De Camargo, O. K., Kerr, E., Rosenbaum, P., Dubey, A., Bostan, C., . . . Cieza, A. (2012). Generic patient-reported outcomes in child health research: a review of conceptual content using World Health Organization definitions. *Developmental Medicine & Child Neurology*, *54*(12), 1085-1095. https://doi.org/https://doi.org/10.1111/j.1469-8749.2012.04393.x

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378. https://doi.org/10.1177/1948550617693063

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191-197. https://doi.org/https://doi.org/10.1016/j.jad.2016.10.019

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1), 72. https://doi.org/10.1186/s12916-015-0325-4

Harding, L. (2001). Children's quality of life assessments: A review of generic and health related quality of life measures completed by children and adolescents. *Clinical Psychology & Psychotherapy*, *8*(2), 79-96. https://doi.org/https://doi.org/10.1002/cpp.275

Hendriks, A. M., Ip, H. F., Nivard, M. G., Finkenauer, C., Van Beijsterveldt, C. E. M., Bartels, M., & Boomsma, D. I. (2020). Content, diagnostic, correlational, and genetic similarities between common measures of childhood aggressive behaviors and related psychiatric traits. *Journal of Child Psychology and Psychiatry*, *61*(12), 1328-1338. https://doi.org/https://doi.org/10.1111/jcpp.13218

Horowitz, J. L., & Garber, J. (2006). The prevention of depressive symptoms in children and adolescents: A meta-analytic review. *J Consult Clin Psychol*, *74*(3), 401-415. https://doi.org/10.1037/0022-006x.74.3.401

Janssens, A., Rogers, M., Thompson Coon, J., Allen, K., Green, C., Jenkinson, C., . . . Morris, C. (2015). A Systematic Review of Generic Multidimensional Patient-Reported Outcome Measures for Children, Part II: Evaluation of Psychometric Performance of English-Language Versions in a General Population. *Value in Health*, *18*(2), 334-345. https://doi.org/https://doi.org/10.1016/j.jval.2015.01.004

Janssens, A., Thompson Coon, J., Rogers, M., Allen, K., Green, C., Jenkinson, C., . . . Morris, C. (2015). A Systematic Review of Generic Multidimensional Patient-Reported Outcome Measures for Children, Part I: Descriptive Characteristics. *Value in Health*, *18*(2), 315-333. https://doi.org/https://doi.org/10.1016/j.jval.2014.12.006

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

Kwan, B., & Rickwood, D. J. (2015). A systematic review of mental health outcome measures for young people aged 12 to 25 years. *BMC Psychiatry*, *15*(1), 279. https://doi.org/10.1186/s12888-015-0664-x

Marsh, H. W. (1994). Sport Motivation Orientations: Beware of Jingle-Jangle Fallacies. *Journal of Sport and Exercise Psychology*, *16*(4), 365-380. https://doi.org/10.1123/jsep.16.4.365

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonson, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2018). *COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) user manual Version 1.0*. COSMIN. https://cosmin.nl/wpcontent/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf

Mullarkey, M. C., & Schleider, J. L. (2021). Embracing Scientific Humility and Complexity: Learning "What Works for Whom" in Youth Psychotherapy Research. *Journal of Clinical Child & Adolescent Psychology*, 1-7. https://doi.org/10.1080/15374416.2021.1929252 Newson, J. J., Hunter, D., & Thiagarajan, T. C. (2020). The Heterogeneity of Mental Health Assessment. *Frontiers in Psychiatry*, *11*(76). https://doi.org/10.3389/fpsyt.2020.00076

Newson, J. J., & Thiagarajan, T. C. (2020). Assessment of Population Well-Being With the Mental Health Quotient (MHQ): Development and Usability Study. *JMIR Ment Health*, *7*(7), e17935. https://doi.org/10.2196/17935

NHS Digital. (2018). *Mental Health of Children and Young People in England, 2017 Summary of key findings*. https://files.digital.nhs.uk/F6/A5706C/MHCYP%202017%20Summary.pdf

Patalay, P., & Fried, E. I. (2020). Editorial Perspective: Prescribing measures: unintended negative consequences of mandating standardized mental health measurement. *Journal of Child Psychology and Psychiatry*, *62*(8). https://doi.org/10.1111/jcpp.13333

Proctor, C., Alex Linley, P., & Maltby, J. (2009). Youth life satisfaction measures: a review. *The Journal of Positive Psychology*, *4*(2), 128-144. https://doi.org/10.1080/17439760802650816

Rajmil, L., Herdman, M., Fernandez de Sanmamed, M.-J., Detmar, S., Bruil, J., Ravens-Sieberer, U., . . . Auquier, P. (2004). Generic health-related quality of life instruments in children and adolescents: a qualitative analysis of content. *Journal of Adolescent Health*, *34*(1), 37-45. https://doi.org/https://doi.org/10.1016/S1054-139X(03)00249-0

Ravens-Sieberer, U., Erhart, M., Wille, N., Wetzel, R., Nickel, J., & Bullinger, M. (2006). Generic Health-Related Quality-of-Life Assessment in Children and Adolescents. *PharmacoEconomics*, *24*(12), 1199-1220. https://doi.org/10.2165/00019053-200624120-00005

Rose, T., Joe, S., Williams, A., Harris, R., Betz, G., & Stewart-Brown, S. (2017). Measuring Mental Wellbeing Among Adolescents: A Systematic Review of Instruments. *Journal of Child and Family Studies*, *26*(9), 2349-2362. https://doi.org/10.1007/s10826-017-0754-0

Rutter, M., & Pickles, A. (2016). Annual Research Review: Threats to the validity of child psychiatry and psychology. *Journal of Child Psychology and Psychiatry*, *57*(3), 398-416. https://doi.org/https://doi.org/10.1111/jcpp.12461

Ryan, R. M., & Deci, E. L. (2001). On Happiness and Human Potentials: A Review of Research on Hedonic and Eudaimonic Well-Being. *Annual Review of Psychology*, *52*(1), 141-166. https://doi.org/10.1146/annurev.psych.52.1.141

Sawyer, S. M., Azzopardi, P. S., Wickremarathne, D., & Patton, G. C. (2018). The age of adolescence. *The Lancet Child & Adolescent Health*, *2*(3), 223-228. https://doi.org/10.1016/S2352-4642(18)30022-1

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, *16*(4), 744-755. https://doi.org/10.1177/1745691620966795

Schmidt, L. J., Garratt, A. M., & Fitzpatrick, R. (2002). Child/parent-assessed population health outcome measures: a structured review. *Child: Care, Health and Development*, *28*(3), 227-237. https://doi.org/https://doi.org/10.1046/j.1365-2214.2002.00266.x

Solans, M., Pane, S., Estrada, M.-D., Serra-Sutton, V., Berra, S., Herdman, M., . . . Rajmil, L. (2008). Health-Related Quality of Life Measurement in Children and Adolescents: A Systematic

Review of Generic and Disease-Specific Instruments. *Value in Health*, *11*(4), 742-764. https://doi.org/https://doi.org/10.1111/j.1524-4733.2007.00293.x

Stevanovic, D., Jafari, P., Knez, R., Franic, T., Atilola, O., Davidovic, N., . . . Lakic, A. (2017). Can we really use available scales for child and adolescent psychopathology across cultures? A systematic review of cross-cultural measurement invariance data. *Transcultural Psychiatry*, *54*(1), 125-152. https://doi.org/10.1177/1363461516689215

Terwee, C. B., Prinsen, C. A. C., Ricci Garotti, M. G., Suman, A., de Vet, H. C. W., & Mokkink, L. B. (2016). The quality of systematic reviews of health-related outcome measurement instruments. *Quality of Life Research*, *25*(4), 767-779. https://doi.org/10.1007/s11136-015-1122-4

Tsang, K. L. V., Wong, P. Y. H., & Lo, S. K. (2012). Assessing psychosocial well-being of adolescents: a systematic review of measuring instruments. *Child: Care, Health and Development*, *38*(5), 629-646. https://doi.org/https://doi.org/10.1111/j.1365-2214.2011.01355.x

Upton, P., Lawford, J., & Eiser, C. (2008). Parent–child agreement across child health-related quality of life instruments: a review of the literature. *Quality of Life Research*, *17*(6), 895. https://doi.org/10.1007/s11136-008-9350-5

Visontay, R., Sunderland, M., Grisham, J., & Slade, T. (2019). Content overlap between youth OCD scales: Heterogeneity among symptoms probed and implications. *Journal of Obsessive-*

Compulsive and Related Disorders, *21*, 6-12. https://doi.org/https://doi.org/10.1016/j.jocrd.2018.10.005

Wolpert, M. (2020). *Funders agree first common metrics for mental health science*. Retrieved 23/02/2022 from https://www.linkedin.com/pulse/funders-agree-first-common-metrics-mentalhealth-science-wolpert

Wolpert, M., Aitken, J., Syrad, H. M. M., Saddington, C., Trustam, E., Bradley, J., . . . Brown, J. (2008). *Review and recommendations for national policy for England for the use of mental health outcome measures with children and young people.* S. a. F. Department for Children. https://www.ucl.ac.uk/evidence-based-practice-unit/sites/evidence-based-practiceunit/files/pub_and_resources_project_reports_review_and_recommendations.pdf

Wolpert, M., & Rutter, H. (2018). Using flawed, uncertain, proximate and sparse (FUPS) data in the context of complexity: learning from the case of child mental health. *BMC Medicine*, *16*(1), 82. https://doi.org/10.1186/s12916-018-1079-6

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1-37. https://doi.org/10.1017/S0140525X20001685

Implications of Paper 4 for Policy Makers and/or Educators and Teachers

Papers 1-3 highlighted challenges in selecting positive and negative mental health measures for assessment, public systems, and judging/selecting interventions. Paper 4 confirmed that a robust evidence base to inform this selection is typically lacking. In addition, measures within domains were not well defined conceptually and many indicators appeared across multiple domains. This makes clear that measure selection should refer to psychometric evidence but also consider the conceptual makeup of items, particularly if multiple measures are deployed. It is likely such conceptual consideration should involve young people (e.g., BeeWell Youth Steering group members, 2021) since Paper 4 suggests this work is often lacking. Where existing measures are selected based on such consultations these should ideally be those with more comprehensive psychometric evidence (e.g., KIDSCREEN).

References

BeeWell Youth Steering group members. (2021). #BeeWell – Measuring wellbeing in secondary schools across Manchester. *what works wellbeing blog*. https://whatworkswellbeing.org/blog/beewell-measuring-wellbeing-in-secondary-schools-across-manchester/

Chapter 5: Discussion

There is a clear need to measure adolescent general mental health via brief self-report methods (see Chapter 1). Briefly, this is because of concerns about this age group in particular (Solmi et al., 2021), problems with data burden and confounding when measuring multiple individual disorders (Deighton et al., 2014), and a lack of clear rationale to preference divergent proxy reports (De Los Reyes et al., 2015). While the prevalence of disorders has tended to be a particular policy concern (Costello, 2015), diagnostic, categorical approaches can have considerable limitations (e.g., reliability Regier et al., 2013; see also Chapter 1). It has also been suggested that positive approaches less focused on symptoms might be more appropriate for early detection in general populations, than assessments of clinical levels of difficulties (Bartels et al., 2013; Greenspoon & Saklofske, 2001; Iasiello & Agteren, 2020).

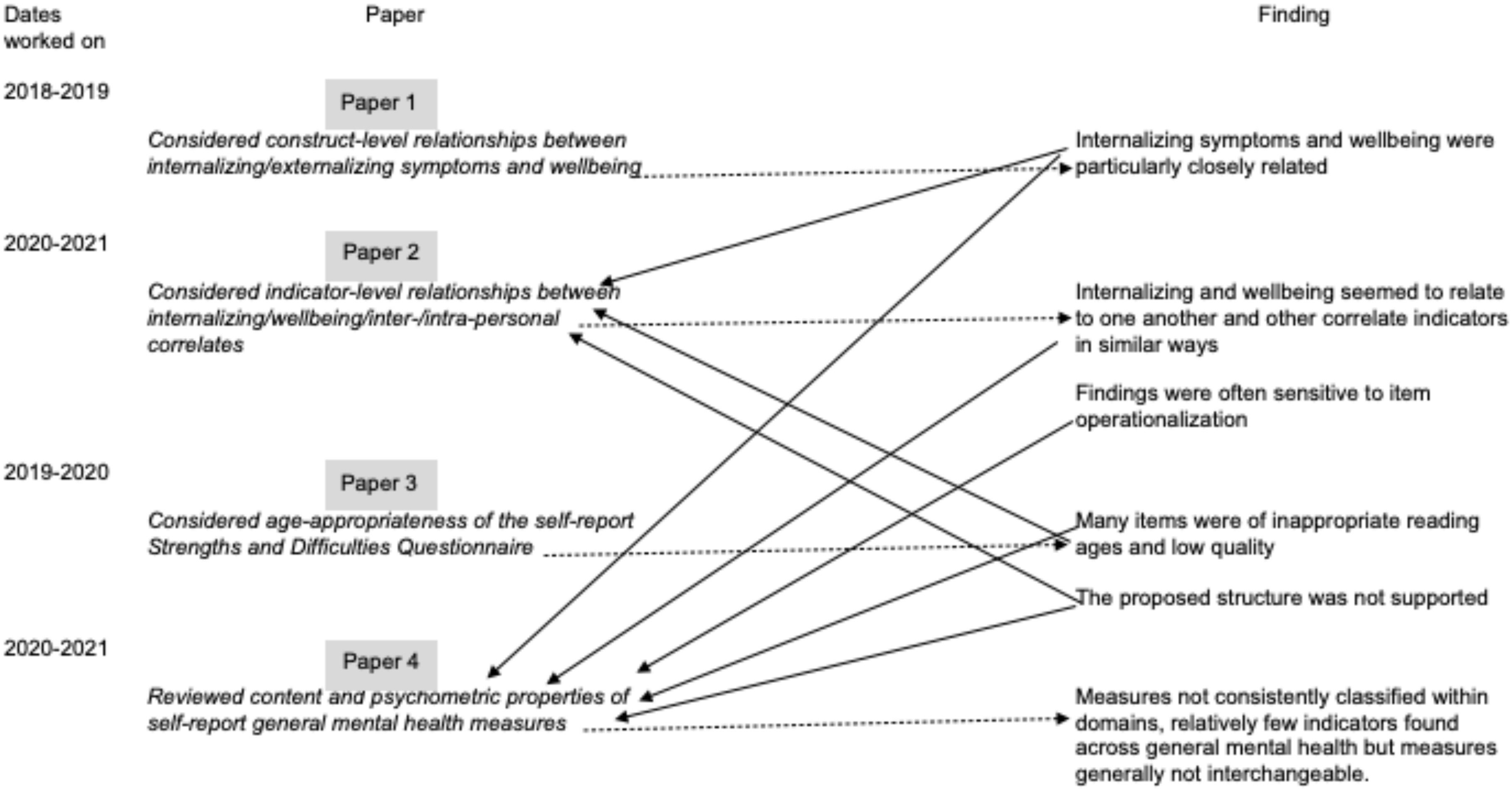
While efforts have been made to include additional positive information, these have tended to be problematic (Moore et al., 2019; see also Chapter 1). Beyond these issues, measure development is typically poor (Flake & Fried, 2020; Flake et al., 2017), and standards appear to be low for measurement with young people (see Chapter 1 and Paper 4). Together these issues suggest there was a major gap in the understanding of general population self-report measurement. While some psychometric evidence for individual scales was available, consideration of empirical and theoretical validity issues was lacking. The current thesis presents initial work in this area, aiming to shed light on robust approaches to measuring general mental health in adolescence.

Summary of Key Findings and How the Papers Informed One Another

As noted in Chapter 4, the papers are presented thematically rather than chronologically. An overview of when papers were worked on and how some of the key findings informed the design of subsequent analyses is shown in Figure 5.1.

Figure 5.1

Overview of How Key Findings That Informed the Design of Subsequent Papers



Dates worked on

Paper

Finding

2018-2019

Paper 1

Considered construct-level relationships between internalizing/externalizing symptoms and wellbeing

Internalizing symptoms and wellbeing were particularly closely related

2020-2021

Paper 2

Considered indicator-level relationships between internalizing/wellbeing/inter-/intra-personal correlates

Internalizing and wellbeing seemed to relate to one another and other correlate indicators in similar ways

Findings were often sensitive to item operationalization

2019-2020

Paper 3

Considered age-appropriateness of the self-report Strengths and Difficulties Questionnaire

Many items were of inappropriate reading ages and low quality

The proposed structure was not supported

2020-2021

Paper 4

Reviewed content and psychometric properties of self-report general mental health measures

Measures not consistently classified within domains, relatively few indicators found across general mental health but measures generally not interchangeable.

Paper 1 considered construct-level models of internalizing and externalizing symptoms and wellbeing. All three constructs were strongly correlated (internalizing/wellbeing and internalizing/externalizing correlations were r = (-).58 in the correlated factors model). For correlations of this magnitude, it is considered prudent to evaluate the extent of multi/unidimensionality (Gignac & Kretzschmar, 2017; Reise et al., 2010). The fact internalizing symptoms and wellbeing were found to be correlated at the same level as internalizing and externalizing symptoms is a crucial finding: Internalizing and externalizing symptoms have sometimes been grouped together and wellbeing argued to be more separate (e.g., Patalay & Fitzsimons, 2016). Moreover, this finding was despite the fact that internalizing and externalizing were measured by a single measure while wellbeing was captured using a distinct instrument (with different implementation properties, e.g., response format). The finding is therefore arguably conservative since common within-measure properties may drive up within-measure correlation (Clifton, 2020; Podsakoff et al., 2003). Indeed, though unidimensionality was not supported across constructs considered, evidence suggested general internalizing explained much of the covariance among item responses in all domains[7]. While more work considering other measures and samples would be important to the generalizability of this finding, it appears congruent with other work which also found strong internalizing/wellbeing correlations (Antaramian et al., 2010; Black et al., 2020; Patalay & Fitzsimons, 2018).

Paper 1 therefore confirmed that positive and negative states could usefully be considered together, and that internalizing symptoms and wellbeing might particularly provide complementary information. Given that much dual-factor research is at least partially motivated by leveraging wellbeing to provide additional insight in general population samples (Bartels et al., 2013; Greenspoon & Saklofske, 2001; Iasiello & Agteren, 2020), this particular relationship between wellbeing and internalizing symptoms is worth highlighting.

Since a general internalizing factor model was not clearly superior to the correlated factor solution in Paper 1, and given calls to consider indicator-level relationships through network approaches, which could equally explain general covariance (van Bork et al., 2017), hierarchical factor models were not used for the consideration of positive and negative mental health in the remainder of the thesis.

[7] Consistent with examples in the literature at the time, ECV was only calculated for the classical, not *S*-1, bifactor model.

In addition, network approaches may afford an additional benefit when considering positive and negative approaches together: One of the problems in the dual-factor literature has been interpreting what it means to be simultaneously high in wellbeing and symptoms of mentally ill health (Iasiello & Agteren, 2020). It may be that this is a heterogeneity problem, as was suggested by the finding of a qualitatively diverse indicator set across general mental health in Paper 4, and as has been described for depression (Fried, 2017; Fried & Nesse, 2015; Fried et al., 2014). Specifically, different individuals might experience different combinations of indicators within each construct or measure. This explanation would allow an individual to experience, for instance, both anxiety and feeling close to others, without the more confusing abstraction of high levels of positive and negative mental health.

Longitudinal network analysis was therefore used in Paper 2 to move beyond factor models and allow consideration of average within-person effects. Partly given the findings in Paper 1, a focus on internalizing symptoms and wellbeing was adopted. In this paper relationships between internalizing, wellbeing and inter/intra-personal correlate indicators were considered over three annual waves. Within-person networks were densely connected, with internalizing, wellbeing and correlates affecting each other in similar ways. While no substantial structural differences were found across gender, other measurement issues were apparent since findings were often sensitive to item operationalizations.

While Papers 1 and 2 focused on relationships between positive and negative mental health constructs and indicators to provide empirical insight into conceptual issues, Papers 3 and 4 considered wider measurement issues. Paper 3 addressed the age appropriateness of the SDQ, a general mental health measure. Though the SDQ is often considered a robust instrument and is widely used and recommended (Vostanis, 2006; Wolpert et al., 2015), the analysis presented here suggests its items were not suitable for subdomain or total scoring, with correlated factor and bifactor models showing poor fit. The measure was found to be invariant across younger and older adolescents (for the flexible ESEM structure), but readability and item quality analysis suggested this could have been because many items were inappropriate for both age groups.

Such measurement issues, revealed across the results and contextual literature for each of the first three papers, made clear a need to review the conceptual and psychometric properties of adolescent general mental health measures (Paper 4). In addition, the content analysis allowed interrogation of the heterogeneity and similarity within and between measures and domains.

Measures were found to be inconsistently classified within domains by reviews. Many measures shared common indicators, though measures within and between domains tended not to be interchangeable in terms of content. The psychometric landscape was also found to be generally poor.

Overall, therefore, the papers of the thesis contributed to two key areas: psychometric and conceptual issues in the measurement of adolescent general mental health. These are discussed in each of the papers. Below findings are brought together from across the papers in relation to the overarching aim (Aim 4), to provide insight into more robust approaches to measuring adolescent general mental health.

Conceptual Issues

Lack of Support for the Separation of Positive and Negative States

A major contribution of the thesis is to our understanding of how positive and negative mental health measurements relate. As discussed in Chapter 1, theoretical and empirical work considering both outcomes have often been problematic, resting on faulty or untested assumptions. A recent review confirms that most adult and adolescent studies considering dual-factor models, relied on model fit comparing unidimensional and correlated factor models, or splitting the sample into the four mental health groups as evidence for the model (Iasiello & Agteren, 2020). In contrast, the current thesis went beyond such methods, considering dimensionality through additional means, employing more appropriate models and considering item content.

Cumulatively, the papers of the thesis suggest a positive/negative dichotomy is unhelpful to furthering understanding, and that prior work has likely overemphasized this. Theoretical and empirical work suggested that domains and indicators within the positive and negative aspects of general mental health considered were not sufficiently dissociated to support the idea that positive mental health is a fundamentally different phenomenon to negative mental health. Statistical findings were often in line with previous work, but the additional considerations made here in terms of analysis and interpretation do somewhat contrast with what has been presented previously. The evidence reviewed and found here suggests that while general mental health is nebulous, a clear boundary between positive and negative mental health may not be the most effective way to organize it as tends to be emphasized in the dual-factor model literature (e.g., Patalay & Fitzsimons, 2016; Putwain et al., 2021; Suldo & Shaffer, 2008).

A number of findings contribute to this conclusion. First, the construct-level correlations in Paper 1 suggested that wellbeing was not more dissociated from internalizing, than internalizing and externalizing were from one another. In addition, as mentioned above, the correlation between internalizing and wellbeing is likely to be more conservative than that between internalizing and externalizing because the latter two constructs were measured via the same instrument. Internalizing and externalizing were also largely worded in the same direction unlike internalizing and wellbeing, which could also lead to relative underestimation of the internalizing/wellbeing correlation compared to that between internalizing/externalizing (Weijters & Baumgartner, 2012). While not all general mental health indicators have positive and negative poles, Paper 4 suggests many can be framed from either perspective (e.g., anxious/relaxed, happy/unhappy, see also the supplementary material for Paper 4 https://osf.io/k7qth/). This suggests that this issue of relative reverse wording needs to be further considered in future work.

Second, while Paper 3 was not explicitly focused on the relationship between positive and negative mental health, the results relating to reverse wording are pertinent to this issue. It was found that reverse worded items loaded most strongly onto the prosocial, or what effectively became the positive factor. This underscores the known issue that reverse wording can influence dimensionality (Weijters & Baumgartner, 2012). In the context of positive and negative mental health, this issue has not been considered, to my knowledge, though it has been suggested that response issues could underlie unexpected patterns of wellbeing and symptom scores (Furlong et al., 2017). While the SDQ is often considered a relatively robust measure that has undergone validation commensurate with standards in the field, and was recommended for domain level scoring, Paper 3 suggested the direction items were worded in had considerable implications for its structure and scoring. Given that most studies considering positive and negative mental health have to rely on *separate* measures, which therefore potentially introduce additional similarity within, and differences between measures (e.g., through response format), this effect warrants attention. I argue that until these potential measurement confounds start to be addressed, arguments emphasizing the dissociation of positive and negative mental health cannot be adequately substantiated.

Third, neither indicators nor measures were specific to constructs theoretically. This was seen through the coding of reviews and content analysis in Paper 4. This lack of consistency for which constructs and indicators belong to which positive and negative domains within general mental health again calls into question holding each outcome separate based on measure or construct names. The review did not discriminate between the direction items were worded in, e.g., happy versus unhappy following other similar work (Newson et al., 2020). Unlike empirical work considering the relationship between positive and negative mental health, Paper 4 was therefore not sensitive to wording or measure effects, concentrating exclusively on theory. It is perhaps surprising that the conceptual overlap of, for instance, happiness and unhappiness needs to be emphasized. However, the work built on the idea that wellbeing and mental health problems are not two ends of the same continuum, has fostered unclear interpretations (see below).

The lack of specificity of indicators to domains was also seen through the steps taken to select items for Paper 2. For example, two items relating to concentration from the SDQ were used to capture internalizing problems consistent with DSM-5 (American Psychiatric Association, 2013), despite the fact it has been argued elsewhere these be considered as indicators of externalizing *not* internalizing difficulties (Goodman et al., 2010). Similarly, *dealing with problems* and *optimism* were both captured in the stress and wellbeing instruments and therefore considered as alternative operationalizations in the multiverse design. This inconsistency and overlap of indicators is typical in mental health measurement (Fried, 2017; Newson et al., 2020), and is therefore perhaps unsurprising. It nevertheless highlights the importance of carefully considering data at the item level, before drawing inferences from statistical analysis conducted at the construct level.

Fourth, when indicator-level analysis, including correlates, was conducted (Paper 2), similar complexity between mental health and correlate indicators was seen, rather than evidence that wellbeing indicators functioned differently to internalizing problem indicators. Some prior work had suggested that different correlates were associated with each of mental health difficulties and wellbeing (Patalay & Fitzsimons, 2016). However, these differences were mostly observed when adolescents reported on wellbeing and their parents reported on mental health difficulties. In a subsequent wave when adolescents self-reported on both outcomes, which predictors were significant was fairly consistent across different positive and negative mental health outcomes (Patalay & Fitzsimons, 2018). The finding of similar relationships to correlates in Paper 2 is therefore consistent with this, given that the focus here was on self-report data.

While more work is needed, the findings of the thesis suggest that researchers and practitioners should not start out with the assumption that a given positive and negative mental health measure provide different substantive information. This is consistent with much of the dual-factor literature which has set out to improve identification by including positive states (e.g., Greenspoon & Saklofske, 2001). This suggests these outcomes are typically considered to provide complementary insights. However, lack of appropriate methods and interpretations have led to problematic inferences in this area that I argue should be dispelled. These are discussed in the following paragraphs. Iasiello and Agteren (2020) describe a consensus of interpretation in their review of dual-factor adolescent and adult studies: Since a correlated factor model fits better than an orthogonal factor or unidimensional model, positive and negative mental health should be considered as "independent and related factors" (p. 5). However, it is not clear what independent means here. Clearly, this is not statistical independence since they are substantially correlated. I argue this finding that the two constructs show some statistical dependence is not surprising since they consider similar indicators, mostly via similar survey methods (Iasiello & Agteren, 2020). Substantive and shared method variance would therefore be expected (Podsakoff et al., 2003). In fact, correlations for convergent validity between measures of the *same* construct (including but not limited to mental health) are not typically so high as to render them statistically interchangeable (Carlson & Herdman, 2012). For instance, in a meta-analysis of self-control measures, self-report questionnaires were found to correlate at r = .50, which the authors considered "strong evidence of convergent validity" (Duckworth & Kern, 2011; p. 265).

It is therefore questionable whether a correlated factors model could be informative about the structure of positive and negative mental states. In fact it has been explicitly recommended that the superior model fit of a correlated factors model over a unidimensional model should not be used to argue the existence of discrete constructs (Gignac & Kretzschmar, 2017). Rather additional steps to consider the extent of multi/unidimensionality should be considered, as was done here. Such additional analysis provided more insight into the distinctness of constructs under study. For instance, the approach highlighted that internalizing indicators particularly tended to share variance with indicators from other domains.

Despite this, superior fit of a correlated factors (over a unidimensional) model, has been fundamental in the field. In a seminal paper Keyes (2005) wrote:

"The current study confirms empirically that mental health and mental illness are not opposite ends of a single continuum; rather, they constitute distinct but correlated axes that suggest that mental health

> *should be viewed as a complete state. Thus, the absence of mental illness does not equal the presence of mental health."* (p. 456)

This paper, and in particular this conclusion, seem to have been highly influential. The first sentence can logically follow from the analysis conducted, a CFA in which the constructs were found to correlate at -.53. However, I argue the second sentence is problematic, and is an example of unjustified inference from a statistical model, which is a problem that has been identified in covariance modeling in general (Fried, 2020). First, if positive and negative mental health are substantially correlated as seems to be the case across multiple studies (e.g., Antaramian et al., 2010; Patalay & Fitzsimons, 2018; Suldo et al., 2011), then each likely provides some information about the other. This again fits with the motivation for much dual-factor work, that the inclusion of positive states should provide more thorough insight and improve screening sensitivity (Greenspoon & Saklofske, 2001).

Second, Keyes' analysis, that of Paper 1, and the factor-model studies reviewed by Iasiello and Agteren (2020), are based on between-subjects data. This means the covariance structure represented describes differences between people. These data are therefore not informative about the presence/absence of states within individuals (Moeller, 2022), which is typically necessary when considering mental states (Fried, 2020). For instance, such models can be insensitive to heterogeneity across individuals (Molenaar, 2004), such that while certain items appear to cluster in between-persons data, this pattern does not apply well to all individuals. This implies that grouping individuals based on categories from measures developed from between-subjects data, e.g., the complete mental health groupings (see Chapter 1 and Paper 1), could have serious consequences.

The approaches in the current thesis improved on these problems. First, though Paper 1 drew on between-subjects data, the analysis could aid sample/population inference such as comparing prevalence, and results are therefore pertinent to the need for improved epidemiological data: Since Paper 1 considered the *extent* of unidimensionality, it provided insight into the implications of analysing positive and negative states together at the group level. The finding that constructs were substantially correlated but not unidimensional demonstrated that straight-forward use of measures together could be problematic, i.e., conclusions about manifest scores of total or individual constructs could be invalid without explicitly modeling additional covariance between constructs.

Second, in Paper 2 indicators rather than constructs were considered, and within-person modeling was afforded through longitudinal data. A key insight that arose from this was that a clear dissociation of correlates was not evident. This is important since disaggregation of within- and between-person effects of the type conducted in Paper 2 allows identification of whether within- and between-person effects are aligned (Moeller, 2022). In Paper 2 within- and between-person networks were *different*, suggesting that covariance patterns among positive and negative mental health indicators in adolescents may not generalize from between to within-person data. Though more work is needed, this would suggest the complete mental health groupings could be further invalidated. Taken together, the current findings highlight potential limitations in what are reasonably widely-accepted conclusions about positive and negative mental health (Iasiello & Agteren, 2020), and the need for more work in this area.

The Importance of Internalizing Constructs and Indicators

Another conceptual theme that can be drawn across various findings in the thesis is the potential value of internalizing indicators. In this context, this includes those explicitly labelled as internalizing symptoms, but also other emotionally-focused indicators, given that internalizing problems are considered to reflect problems in this area (Graber & Sontag, 2009), and inconsistencies highlighted in Papers 2 and 4. These indicators were found to be empirically important and strongly represented across measures. These findings are consistent with the fact adolescence appears to be a sensitive time for the development of these problems (NHS Digital, 2018; Rapee et al., 2019). This sensitivity could explain why scale developers seem particularly keen to capture these experiences in measures (Paper 4), and why they strongly covary with others (Papers 1 and 2). Specific findings in relation to the importance of emotional indicators are discussed below.

In Paper 1 the most appropriate hierarchical model had a general internalizing distress factor which predicted meaningful variance in all internalizing, wellbeing, and externalizing items. The simpler, and therefore generally preferable, correlated factors model represented this covariance structure via strong correlations between all constructs, but particularly between internalizing and each of the others. For either model it is clear that the internalizing items were strongly related to those from both other domains. While the relationship between internalizing, wellbeing and externalizing likely varies across individuals (Molenaar, 2004), and measures, at a whole-sample level, measuring emotional/internalizing states may be an efficient way to capture initial insight into general mental health.

As noted in Paper 1, results could have been affected by the fact it may be preferable to use proxy informants for externalizing problems. Nevertheless, the finding that when using self-report methods, which as argued in Chapter 1 are beneficial for several reasons, internalizing/emotional indicators tended to share variance with others remains important. This shared variance may be indicative that self-report internalizing indicators can provide some insight into other states. However, more work to explore this is needed. Furthermore, work is needed to consider how to efficiently and validity estimate externalizing problems in large samples, and if this is possible via self-report, since prior work has shown that adolescents experiencing both internalizing and externalizing disorders might represent a particularly vulnerable subgroup (McElroy et al., 2017).

In Paper 2 only internalizing and wellbeing indicators were selected, given the evidence from Paper 1, and similar work (Antaramian et al., 2010; Black et al., 2020; Patalay & Fitzsimons, 2018), that these may provide complementary insight (particularly covary), and the need to focus on internalizing problems, given particularly high prevalence in adolescence (NHS Digital, 2018). Therefore, since they were excluded, this paper was *not* informative about the relative importance of externalizing symptoms. As mentioned above, work to provide methods into these experiences is perhaps particularly needed. However, the relative centrality of internalizing symptom, wellbeing and correlate indicators was calculated. While some results were sensitive to specifications, of the four indicators that were more consistently influential, two were unequivocally emotional: *unhappy* and *worry*. The other two were *think clearly* and (dealing with) *stress*. Thinking clearly is also considered a diagnostic criterion for depression and anxiety (American Psychiatric Association, 2013), consistent with the empirical importance of internalizing symptom indicators found in Paper 1 and the fact adolescence is considered a sensitive period for the development of these disorders (Rapee et al., 2019). Findings from Paper 2 therefore provide tentative additional evidence for the importance of internalizing/emotional indicators (see also section below on key indicators).

As briefly mentioned, the relative importance of emotional/internalizing indicators was also suggested in Paper 4: Most items were coded as tapping emotional themes, and the most frequent indicators across measures could also be considered under this category: (un)*happiness*, *enjoyment*, *self-worth*, *fear/worry*, *loneliness/withdrawal*. However, some caution is needed. Findings could be explained by the fact behavioural and cognitive indicators tended to be limited to symptom measures, and were therefore less prevalent because they only appeared in one of the domains under study. The relatively low representation of these indicator types could also reflect the review's focus on self-report measures. Though relatively little work has considered the validity of self- versus proxy-informants for adolescent mental health symptoms, there is some consensus that self-reports may be less preferable for externalizing outcomes, while for internalizing these are often considered the most valid (De Los Reyes et al., 2015; Humphrey & Wigelsworth, 2016). Work to clarify the validity of self-report externalizing indicators and their conceptual relevance is therefore needed before firm conclusions can be drawn about the relative theoretical importance of emotional/internalizing indicators over and above externalizing symptoms.

Together these findings suggest that emotional/internalizing states may be key when measuring general mental health in adolescence via self-report. However, more work is needed, particularly considering within-person models (as used in Paper 2), and drawing on carefully developed measures, particularly for externalizing symptoms.

Key Indicators

In addition to insight about the broad importance of emotional and internalizing states, the item/indicator-level approach adopted across the thesis, allowed identification of potentially key indicators. While, as argued in Paper 4, this must be informed by work with young people, findings from the current thesis may offer a starting point and can already provide insight into existing approaches.

For the relative importance of indicators, the papers of the thesis allowed consideration of factor loadings (Paper 1), network centrality (Paper 2), readability and item quality (Paper 3), and representation across measures (Paper 4). A summary of key indicators, evidence supporting them and the implications is available in Table 5.1. Two indicators, *happiness* and *worry*, were identified across each of these four considerations, while certain others showed some potential importance but were less consistently flagged across papers. While those that were uniformly identified are likely important to include in measures and studies, the differences between methods used across papers, and challenges with the quality of measurement discussed in Chapter 1 and Paper 4, mean that those identified less are not necessarily less important. In addition, the focus of the thesis was on measurement in general population samples which might mean certain indicators associated with clinical mental health problems would have too little variability to show particular covariance, or they may even have been excluded from measures during development for this reason. Despite this, such indicators may still be important for screening purposes. Therefore, papers such as those presented in the current thesis cannot be the only groundwork for improving conceptualization and measurement. Work with stakeholders is still needed to address issues such as these.

Table 5.1

Key Indicators, Supporting Evidence and Implications

Indicator	Evidence	Theoretical and empirical?	Implications
Happiness	High loading (P1) Central (P2) Common (P4)	yes	Indicator should be included
Worry	High loading (P1) Central (P2) Age-appropriate (P3) Common (P4)	yes	Indicator should be included, SDQ item is appropriate
Loneliness	High loading (P1) Common (P4)	yes	Conceptual issues to be clarified
Enjoyment	High loading (P1) Common (P4)	yes	Work considering sensitivity needed
Self-worth	High loading (P1) Common (P4)	yes	Conceptual issues to be clarified
Dealing with stress	Central (P2)	no	Conceptual issues to be clarified

Think clearly	Central (P2)	no	Conceptual issues to be clarified
Anger/ temper	High loading (P1)	no	Informant issues to be clarified

Note. P1 = Paper 1; P2 = Paper 2; P3 = Paper 3; P4 = Paper 4; SDQ = Strengths and Difficulties Questionnaire.

Some of the inconsistency across papers could also reflect a lack of consensus for which indicators belong to mental health or to proximal constructs (see also Paper 4). The following indicators could be considered under this category: *loneliness, dealing with stress, enjoyment and self-worth*. In addition, two further indicators that were flagged but less consistently so, *anger/temper* and *think clearly*, likely would be considered by most to be part of mental health. However, these may have been less well represented, since only Paper 1 included externalizing problems, and only Paper 2 included eudaimonic wellbeing. Each of these indicators are discussed below.

Happiness and Worry*.* *Worry*, including school worry, and (un)*happiness* had among the strongest loadings on the internalizing factors in the S-1$_{\text{Internalizing}}$ and correlated factors models (λ = .64-.70, see Figures 2 and 5 in Paper 1). These items were also among the nodes with more consistent high strength centrality, which is congruent with high loadings in a factor model (Hallquist et al., 2019). The findings from Paper 1 and Paper 2 therefore both suggest that *happiness* and *worry* indicators might particularly share variance with other elements of mental health (and inter/intrapersonal correlates in the case of Paper 2). The analogous nature of strength centrality and factor loadings, means this effect can be interpreted as robust to the different measures and age ranges used in the two papers. These items were also shown to be invariant across gender (Paper 1) and age (Paper 3). However, the related item about crying in Paper 1 was not invariant across gender, suggesting continued careful operationalization and work with stake-holders is needed to ensure validity. Future work could also impose stricter constraints and build on the exploratory work presented here. For instance, it could be important to analyse whether endorsements of these items were predictive of clinical need or prognosis, or specify formal models considering how these symptoms influence others (Haslbeck et al., 2021).

The theoretical importance of *happy* and *worry* was explored in Paper 4. Except for *worry* in life satisfaction, these two indicators appeared across all of symptom, wellbeing, quality of life, affect and life satisfaction domains, and in almost all measures. While this paper suggested conceptualization was generally immature in adolescent general mental health, it was clear *happy* and *worry* are near universally considered important, at least by measure developers. One of the recommendations from this paper was to consider particular operationalizations of such important indicators that had evidence of consultation with young people. This would ideally ensure conceptual importance as well as age appropriateness and possible differences in interpretation between genders. Evaluating the former was beyond the scope of the current thesis, but age-appropriateness was considered in some detail in Paper 3.

The exact *worry* and (un)*happiness* items used in Paper 2, since they came from the SDQ, were subjected to readability and item quality analysis in Paper 3. “I worry a lot” was the *only* emotional symptoms item to score positively for item quality, and had a substantially lower reading age (5.41) than the other emotional symptoms items. This item is simply worded, while the other emotional symptoms items (including the (un)*happiness* item) all reference multiple experiences or symptoms. Given evidence of its age appropriateness, theoretical importance, and ability to capture constructs or predict other indicators, this *worry* item provides an example of one that should be taken forward when considering new measures and analyses.

In sum, the combined theoretical and empirical support for *happiness* and *worry* suggests these are likely key for inclusion, though more work to find appropriate operationalizations should remain a priority. Interestingly, while *happiness* and *worry* may be considered archetypical symptoms of depression and anxiety, or cornerstones of the internalizing spectrum (Graber & Sontag, 2009), Paper 4 suggests they are also often considered beyond disorder criteria. Their strong relationship to the general internalizing factor in Paper 1 and centrality in Paper 2, also suggests their empirical importance for a range of general mental health constructs and indicators. Therefore, though there are good reasons to move beyond disorders (see Chapter 1), moving to exclusively strengths-based approaches, could be problematic (Humphrey & Wigelsworth, 2016). These could miss important experiences, but it is also likely that though they are labelled as strengths-based, some measures and domains in fact capture symptoms (an example of the jangle fallacy).

Loneliness, Enjoyment, Self-Worth, Dealing with Stress and Think Clearly. These indicators showed inconsistent results across the papers of the thesis but were identified in at least one. Notably, they were also not *represented* consistently (in available measures) across papers, reflecting the variable conceptualization identified in Paper 4. The indicator codes in Paper 4 were also as broad as possible to avoid false positive differences between measures (Newson et al., 2020). Coding was also contingent on the specific measures included, such that where it was not clear that a pair were different, they were grouped together, again following other work and to be conservative (Fried, 2017). While broad coding was appropriate for the design of Paper 4, this approach may be more limited when considering the question of key indicators: The fact that conceptualization is immature, and that limited work to ensure content validity has been conducted, means the boundaries of indicators are inherently poorly defined. Comparison of common indicators in Paper 4 to empirically important indicators in the other papers should therefore be considered in light of this.

Loneliness was flagged as important through virtue of having one of the highest loadings in Paper 1 on the internalizing factors for the $S\text{-}1_{\text{Internalizing}}$ and correlated factors models ($\lambda = .68$ for both, see Figures 2 and 5). It also appeared in the majority of measures, and all but the life satisfaction domain, in Paper 4, through the loneliness/withdrawal code. This focused on a relatively wide range of emotional aspects of social connections and isolation, including feeling loved, and feeling able to trust others (see supplementary material for Paper 4 https://osf.io/k7qth/). On the other hand, the relationship satisfaction code covered mostly items from life satisfaction measures, which tended not to focus so explicitly on the emotional aspect. This means that social experiences were represented across all domains, though arguably with variable affective content. While these groupings are somewhat broad when considering key indicators, given the aims of the paper, the distinction made is somewhat consistent with the classical definition of loneliness as painful and arising from a deficit in social interactions compared to those desired by an individual (Cole et al., 2021; Perlman & Peplau, 1984).

While *loneliness* and *withdrawal* were grouped together in Paper 4, only (social) *withdrawal* was available in the dataset used for Paper 2, and this was not among the most central items for any network or condition. This discrepancy suggests that though differences between *loneliness* and *withdrawal* were not clearly defined when looking at the group of measures included and erring on the side of caution in Paper 4, important differences may exist. This is consistent with theory that loneliness is particularly salient for mental health when there is a failure to re-establish social connection (Qualter et al., 2015). In fact, in Paper 2, the *withdrawal* item was unexpectedly related to school support, such that more *withdrawal* was associated with greater feelings of being cared about by an adult at school. As discussed in the paper, it is therefore likely that this particular operationalization of *withdrawal* (which indeed implied a focus on peer relationships), captured a specific social experience rather than an experience related to loneliness, and in fact at least some of those endorsing this seemed to be garnering social support elsewhere. Alternatively, the lower quality identified in Paper 3 of the SDQ *withdrawal* item may have introduced noise which attenuated effects for this node. Conversely, the item in Paper 1, which had been developed in consultation with young people, is much simpler, “I feel lonely”, similar to the SDQ worry item and consistent with other similar work (Sydney & Pyle, 2018). While other factors, such as gender invariance remain important, it is to be expected that clearly interpretable items would be more sensitive, and as argued above should therefore be preferred.

Of the remaining indicators in this category, *enjoyment* and *self-worth* were identified as appearing across most measures in Paper 4. The wellbeing measure in Paper 1, the CORS, did capture *enjoyment* which was the code used for life satisfaction items (and also included interest/pleasure in activities, see supplementary material for Paper 4 https://osf.io/k7qth/). However, loadings of CORS items are arguably not as informative about their importance for general mental health as the internalizing items. This is because of the wording similarity between items, and because the internalizing items could be considered from the perspective of the general internalizing distress factor. Nevertheless, the item, “how is everything going?”, which might be akin to a general life satisfaction item, had a substantially higher factor loading compared to the other items, suggesting this best captured the shared variance among indicators ($\lambda = .73$ and .88 on the wellbeing factor for the correlated and $S\text{-}1_{\text{internalizing}}$ models respectively), compared to specific satisfaction-type items about e.g., school or family. More work, including drawing on within-person data, would be needed to determine the appropriateness of general versus specific life satisfaction items and their importance for general mental health. This is particularly the case since it has been argued domain-specific approaches to life satisfaction are needed for sensitivity (Antaramian et al., 2008), and this is not something that was tested in the current thesis.

The *self-worth* code included satisfaction-type items about 'me', as well as items about confidence (see supplementary material for Paper 4 https://osf.io/k7qth/). While self-esteem measures were not included in any of the papers of the thesis, the CORS *Me* item and SDQ *nervous/lose confidence* item[8] would fit under the broad coding used in Paper 4. As before, the loadings for CORS are arguably less informative than for the internalizing items in Paper 1, though this item did show the second highest loading (λ = .67 and .47 on the wellbeing factor for the correlated and $S\text{-}1_{\text{internalizing}}$ models, respectively). Similar to the *loneliness/withdrawal* issue, the *nervous/confidence* item did not show higher centrality for any model, and again, it was an item that was shown to contain multiple statements and have lower quality in Paper 3. As before, therefore, the lack of clarity in conceptualization, and item problems, mean the importance of self-worth, and likely sub-indicators within this, could perhaps have been obscured or overemphasized. Paper 1 also highlighted inconsistencies in the prior literature for self-esteem, with some including this as part of complete mental health models (St Clair et al., 2017), many not considering it (e.g., Suldo & Shaffer, 2008), and others explicitly treating it as an external predictor (Greenspoon & Saklofske, 2001). Consideration of the theoretical role of *self-worth* for adolescent general mental health is therefore particularly needed.

Finally, *dealing with stress* and *think clearly* were also highlighted as a more consistently central items in Paper 2, but *cope with problems* and *under strain* (either of which might encapsulate stress) appeared in relatively few measures/domains in Paper 4. Analogous stress and thought items were not available in the measures used in Papers 1 and 3, limiting comparisons. In terms of stress, there was also an additional node in Paper 2, for which two operationalizations were available, *dealing with problems.* This node was sometimes among the most central but varied substantially. As alluded to above, this poor representation in Paper 4 could reflect that these might often be considered proximal indicators (Fritz et al., 2018), and indeed they were often not drawn from one of the explicit mental health measures in Paper 2 (one of the problems items and *think clearly* came from SWEMWBS). Nevertheless, the analysis in Paper 4 suggests this kind of indicator might *sometimes* be considered part of general mental health, and the results from Paper 2 suggest they could be important indicators to capture.

[8] This SDQ item was included in Paper 4 and was double coded as *worry* and *self-worth*, given both are covered in this single item. The other mental health measures used in the empirical papers, CORS, M&MS, and SWEMWBS were not present in included reviews for Paper 4, though the related measures ORS and WEMWBS were. This could reflect the quality of the reviews or psychometric studies underpinning these measures.

These inconsistently identified indicators should be a focus for future conceptualization work since their boundaries and importance remain unclear. Researchers and practitioners should also not assume that indicators not included in measures are not important. While this may seem obvious, the tendency to conflate measurements with constructs, despite advice against this, is common (Yarkoni, 2020).

Anger/Temper*.* In Paper 1, high loadings were found for the anger and temper items, including on the general internalizing distress factor λ = .55, .50 (see Figures 2 and 5). These externalizing indicators, should therefore also be further investigated in terms of their appropriateness for general mental health measurement. The S-$1_{\text{internalizing}}$ model suggests two relatively substantial portions of variance were attributable to each externalizing item, those of the general internalizing distress and specific externalizing factors. As discussed in Paper 1, it is possible these could relate to the elements related to distress and pure behavioural problems respectively. However, specific work to test this hypothesis would be needed, as well as to improve and ascertain the appropriateness of measuring these via self-report (as also argued above).

Some insight into these indicators is also available from Paper 3. The SDQ has a single anger/temper item along with other conduct problems. In this paper, for the ESEM solution, the highest loading item on the conduct problems factor was “I fight a lot. I can make other people do what I want” (see Appendix 3). This item was the only one for this scale to not show meaningful crossloadings, partly explaining its stronger relationship to the factor compared to the others. Conversely, the anger/temper item had similar loadings on both the conduct and emotional symptoms factor. The five conduct problems items had relatively low reading ages compared to others (9.26 for the subscale), but all but one scored negatively for quality due to reverse wording or multiple statements. The loadings should therefore be considered in light of this, namely that crossloadings could also reflect participant confusion, thus contributing to an unclear structure. Nevertheless, the relationship of anger/temper to the emotional problems subscale fits with the high general internalizing distress loading in Paper 1, the fact these are explicitly emotional experiences (rather than behavioural), and related indicators such as irritability can be considered criteria for depression (Fried, 2017). It may therefore be that anger/temper are particularly useful when considering the emotional aspect of general mental health. This is not to say that other behavioural aspects should be discounted but again more work is needed to consider this.

As noted in the above section on internalizing/emotional indicators as a whole, the findings for these externalizing indicators are perhaps particularly affected by the design of the papers. Briefly, Paper 1 included externalizing symptoms but different insight would likely have been afforded had proxy informants been used. Externalizing symptoms were excluded from Paper 2 based on the findings of Paper 1 and the need to focus particularly on emotional mental health (NHS Digital, 2018). Finally, externalizing symptoms may have been underrepresented in Paper 4 due to the focus on self-report and likely exclusion of this from domains other than symptoms. Future work should therefore consider anger/temper indicators, since there was some evidence that these are valuable. However, other behavioural indicators should also be examined since these may have been raised less in the exploratory work of the current thesis due to analytical decisions.

Summary of Considerations for Key Indicators*.* Some have argued that selecting measures, and therefore indicators, may need to be done in a context-specific way, particularly given inconsistencies in the current field (Patalay & Fried, 2020). However, the work of the current thesis, which has started to provide insight into key indicators, could inform harmonized approaches. These are necessary for comparison between studies (Krause et al., 2021), something that seems particularly important for broad approaches to mental health in general population samples. In addition, the combination of quantitative and qualitative approaches in the current thesis made clear that indicators and items must be selected based on a range of evidence: Conceptualization must be clarified with key stake-holders, age-appropriateness must be considered, and empirical covariance with other mental health and relevant correlate indicators and constructs should be evaluated.

These are all key steps in improving the quality of data collected which in turn could help move away from the correspondence theory to the coherence of truth (Kendler, 2016). As described in Chapter 1, the correspondence theory in which measure validity is judged against a single criterion (e.g., diagnostic framework) is inappropriate for mental health data. On the other hand, the coherence account requires garnering broader insight into phenomena so that validity can be judged based on how well measurements fit with what we know overall. Until this more comprehensive insight from experts and young people is gained, I argue this coherence account is out of reach. The current thesis presents some initial work in this area, and has identified a constellation of indicators that should be particularly attended to for the conceptualization of general mental health. Having provided insight into the field as it currently is, putative indicators and domains could be used as starting points in such consultations and future analyses.

Summary of Conceptual Issues

The discussion above makes clear that the current thesis contributed to knowledge about conceptual issues in measurement of general mental health in adolescents from multiple angles. A clear conclusion that challenges much prior literature, is the lack of evidence for considering positive and negative states as distinct. That is not to say that individual constructs and indicators within general mental health do not have a role to play, but rather that critical consideration of conceptual issues has often been lacking and must be included going forward. The papers of the thesis set examples and groundwork for approaches to doing so.

The two other contributions to conceptual knowledge are more tentative since little previous work has tried to uncover which aspects of general mental health are important when agnostic to theoretical domains. Put another way, work has tended to be somewhat siloed, whereas the initial findings presented here, suggest much could be gained through concerted efforts to bring together expertise across domains of general mental health. While some work on individual measures has taken a more bottom-up approach through consultation with young people and experts, this has happened relatively infrequently (see below and Paper 4), and has not brought work from across domains together. Future work to consider key states and indicators must therefore draw on consultation with stakeholders. Nevertheless, the potential importance of emotional indicators in general, and *happiness*, *worry*, *loneliness*, *self-worth*, *enjoyment*, *anger/temper* and *dealing with stress* in particular was suggested by multiple methods and findings. Much more work is needed to explore the sensitivity, conceptual and age appropriateness of these indicators, as well as the best strategies for their joint measurement.

Psychometric Issues

Beyond the conceptual issues outlined above, various psychometric findings spanned the papers of the current thesis. Overall, Paper 4 found psychometric properties to be poor and this was seen in various details in the other papers. Common psychometric themes across the papers of the thesis are discussed below.

A key psychometric issue across the field (Paper 4) was a lack of content validity, for which to score positively, measures had to have evidence of young people’s involvement (Terwee et al., 2007). This is notable since other psychometric properties are considered to be unclear where content validity is lacking (Mokkink et al., 2018). As noted in Chapter 3, though secondary analysis was used in the thesis to capitalize on available datasets, relatively few had appropriate positive *and* negative dimensional measures. Of those used, only Paper 1 was able to draw on measures that had any evidence of consultation with young people, further suggesting there is a major deficit in this area.

This issue of a lack of consultation with young people could contribute to the readability and item quality problems seen in Paper 3. It could also underlie some of the conceptual inconsistency seen in Paper 4, and differential key indicator findings between papers for some indicators (see above). Asking young people which feelings and experiences are important would likely increase clarity. This is vital given the findings from Papers 2 and 4: Both papers found different measures/domains to contain common indicators. However, when these were varied as alternative operationalizations in Paper 2, results also varied. This suggests that items were interpreted differently by young people, and as discussed above, conceptual and age-appropriateness issues could both play a role in this.

Paper 4 also found structural validity was often not evaluated and, where positive, was limited either by this lack of content validity, or inconsistent evidence of fit. Structural problems were also seen in the other papers. The structural validity of the measures used in Paper 1 was considered for each individually, and modifications to the published structures were necessary to meet standard fit criteria for both. While the modifications in Paper 1 were relatively minor, the ESEM solution needed to overcome structural problems in the SDQ (Paper 3), represented a much more radical shift away from the measure's reported properties. Though this is a widely used and accepted measure (Vostanis, 2006; Wolpert et al., 2015), the structural findings in Paper 3 were consistent with other work (Goodman et al., 2010; Percy et al., 2008), suggesting this was not a sampling issue.

Consistent with these fundamental content and structural problems, and as seemed likely from the issues considered in Chapter 1, Paper 4 found that a lack of rigorous psychometric development *characterised* the field. While a lack of extensive psychometric studies does not preclude desirable measurement properties, perhaps particularly if based on sound theory or careful age-appropriate design, this approach is irresponsible and potentially a major waste of resources. The comprehensive approach taken in Paper 4, in which similar domains were considered together and rated via the COSMIN criteria provided important insights. For instance, though other domains such as wellbeing and perhaps particularly quality of life, might appear to have less problematic development practices than symptom measures (see Paper 1 and Paper 4), partly since they do not rely on problematic nosology, all domains showed significant psychometric issues. Moreover, the analysis presented in Paper 3 suggests far-reaching consequences can arise from lax practices: SDQ total and subscale scores should be treated with extreme caution as these are not empirically supported. As has been urged in psychology more generally (Flake et al., 2017), these findings therefore serve as a warning not to blindly trust measures which have not undergone thorough investigation.

Further demonstration of this was seen for reliability. While basic reliability statistics are often relied on as evidence of measures' psychometric quality, this is a poor practice (Clifton, 2020; Flake et al., 2017). In addition, the most commonly used internal consistency metric, alpha, suffers from numerous problems including the unrealistic assumption that items should be equally weighted and particularly deflates results for short scales (McNeish, 2018). The findings of the current thesis provide an illustration of this and its implications. In Paper 4 internal consistency was available for every included measure (see supplementary material for Paper 4 https://osf.io/k7qth/). However, most were rated as indeterminate according to the COSMIN system since they lacked structural validity evidence (which is required to score any higher). Considering the SDQ again, it passed standard thresholds for several internal consistency coefficients in the sample considered in Paper 3, but its structure was clearly not supported, and ESEM analysis demonstrated items sometimes more strongly tapped other domains than their labelled subscale. This demonstrates that relying on internal consistency, rather than more comprehensive psychometric analysis, can result in substantial threats to the robustness of studies and monitoring.

Beyond the appropriateness of reliability, structural measurement models can also be used to perform measurement invariance analyses. Paper 4 suggests this has not typically been considered, though it is vital to comparing groups (Millsap, 2012). Papers 1, and 3 therefore contributed to the field by conducting such analyses. Measurement invariance between adolescents who had ever been in receipt of free school meals versus those who had not, in Paper 1 was supported, and only small, possible negligible, differences were seen between girls and boys. Invariance across age groups was also supported for the SDQ in Paper 3. While measurement invariance in a strict sense was not considered in Paper 2, the similarity of relationships between mental health and correlate indicators was considered and only very small differences were found between girls and boys. These findings are encouraging, though measurement invariance must be considered alongside wider validity issues.

For instance, as discussed in Paper 3, given the readability and item quality issues found, invariance could hold because *neither* group interpreted items as intended. This again highlights the need to attend to fundamental aspects of measure development, notably content validity.

Summary of Psychometric Issues

The findings of the current thesis are unequivocal: The psychometric properties of adolescent general mental health measures are understudied and often poor. The problems highlighted above at the conceptual level are likely key to rectifying this, though advanced modeling (such as that used in the current thesis) has also been underutilised, with heuristics such as coefficient alpha relied on.

Combining Advanced Quantitative and Conceptual Approaches to Psychometrics

The papers of the thesis not only provided insight into quantitative issues, but provided examples of approaches to start to mitigate potential problems and further understanding. For instance, in Paper 3, quantitative psychometric results were considered alongside possible explanations for psychometric issues (readability and item quality), in the absence of adequate development procedures. Similarly, while Paper 3 made clear the SDQ's published structure was problematic, Paper 2 took an alternative approach (network modeling), and provides an example of the type of analysis that could be used to try and maximize use of the data. Crucially, Paper 2 was not just a move away from sum-scoring or the proposed factor structure, but detailed considerations of how items were selected and the rationale for each were also included (see Paper 2 and its supplementary material https://osf.io/rxv5q/). The approach in the current thesis in which issues beyond model fit were considered (see Chapter 3), therefore enriched understanding. This is important since it seems clear standards are generally low in adolescent general mental health measurement (see Chapter 1 and Paper 4). Content validity evidence is particularly lacking. Given its fundamental importance (Mokkink et al., 2018), this absence arguably makes attention to issues beyond model fit even more important.

While Paper 4 primarily provided an overview of the field, it also introduced the idea of jointly considering statistical and conceptual consistency, which provides insight into the applicability of statistical homogeneity for a given conceptual space. As discussed in Paper 4, the HS (conceptual homogeneity/statistical consistency) scoring approach is rather basic, and this was necessary given the scope of the paper as a whole. Nevertheless, the question of appropriateness of sum scoring has been raised for other domains of mental health, notably depression, because of indicator heterogeneity, and poor reliability of diagnoses (Fried, 2017; Fried & Nesse, 2015; Fried et al., 2014).

The current thesis highlights inconsistencies in general mental health measurement in adolescence, at construct and indicator levels. I argue that this, in combination with poor development practices leading to problems such as inappropriate reading ages and structural invalidity, suggest that these inconsistencies are likely not solved by scoring and combining existing domains (e.g., wellbeing and mental ill-health). Rather, more work is needed to bring together theoretical and quantitative psychometric issues and as argued above, draw on expertise from across approaches to general mental health. In the interim, the findings presented here, confirm that general mental health is, as expected, a diverse construct. Since there is a need to capture it in a brief way, more work is needed to discover sensitive indicators for general populations and evaluate the cost/benefit implications of sum-scoring.

Strengths and Limitations

Each paper describes specific strengths and limitations for each study, which are not all reiterated here. Instead, a few overarching issues and common themes are highlighted.

As a whole, the thesis benefits from drawing on several large samples (or subsamples), and considering issues in several measures of mental health difficulties and wellbeing. The use of secondary data also provided insight into measures that had been explicitly selected to estimate general mental health (consistent with the aims of the thesis). Conclusions can therefore be compared across measures and samples in some cases. For instance, the discussion above of key indicators brings together several congruent insights from across the papers, and also highlights some areas where there is a need for clarification.

In addition, robust methodological and open science practices were a key focus of the thesis. Statistical procedures and systematic review methods were reviewed for each paper and the most appropriate and robust were selected based on best practice recommendations (e.g., estimation procedures and PRISMA, see also Chapter 3). The difficulty involved in robustly applying psychometric methods (Borsboom, 2006), mean studies often rely on basic heuristics like coefficient alpha (Flake et al., 2017), or make unrealistic inferences based on more complex psychometric modeling (Haeffel et al., 2021). On the other hand, the current thesis took account of a wide range of methodological issues (e.g., age-appropriateness and within- versus between-person effects), in a fast-paced field, and aimed to report these transparently to aid interpretability, replicability, and future work (Epskamp, 2019). Papers 2 and 4 were pre-printed, and materials were openly presented alongside publications. The data on which this thesis is based are not currently openly available. However, for Paper 2 synthetic data were provided, which allows others to run the code. This was an important consideration, given the novelty and complexity of the method. Providing early examples of new methods was in itself a major strength of the thesis, with among the very earliest applications of *S*-1 and panel network models presented for others to draw on. Similarly, the introduction of a framework to consider conceptual and statistical consistency together in Paper 4 could have implications within and beyond adolescent mental health.

Despite these strengths, certain limitations must be acknowledged. First, Papers 1-3 could be limited by relying on non-probability samples, and findings from Papers 1 and 2 might not generalize to other measures, samples or populations. However, given the approach to reporting across the thesis, and particularly the multiverse design of Paper 2, generalization across measures does not limit the conclusions drawn. Indeed, sensitivity to measurement operationalization is a key focus of the thesis. Second, as is usual in secondary analysis (see Chapter 3), a-priori power was not considered and it is therefore possible that Papers 1-3 could face issues in this area. While post-hoc Monte Carlo simulation studies could have been used for Papers 1 and 3 (power methods for network models are lacking; Aalbers et al., 2019), these would not have been useful in informing preprocessing decisions. Most importantly, the likely considerable power afforded by the large sample sizes included, and consistency/transparency of pre-processing steps used were considered as key strengths (see also Chapter 3). Sample sizes also met basic rule-of-thumb guidelines (Kline, 2015). Where sample size was likely to create oversensitivity for chi-square difference testing and Type I error, and this was used to judge models, results were also interpreted cautiously, i.e., in line with partial measurement invariance recommendations (Paper 1). In Papers 2 and 3 model comparison was not based on chi-square difference testing for this reason and instead used AIC/BIC and CFI difference respectively.

Third, the quality problems with the SDQ highlighted in Paper 3, could have implications for the validity of Paper 2. However, as described in Chapter 1, Paper 3, and Paper 4, psychometric development has been generally poor in the field, and other secondary datasets were not suitable. In addition, since Paper 3 was actually published before Paper 2 (see Chapter 4 and Figure 5.1), item choices were informed by findings from Paper 3: more complex items were rejected where possible, i.e., when alternatives were available in the dataset for the same indicator (see also supplementary material for Paper 2 https://osf.io/rxv5q/).

Fourth, wellbeing and mental health problems were always measured via different instruments with different implementation properties (e.g., response format), as is the case across the greater literature (see Chapter 1). The consistency of these properties within measures, and differences between measures, as well as the reverse wording of these constructs relative to one another, could have downwardly biased associations across domains. However, Paper 4 suggested that measures considering both aspects in a single measure do not exist, or at least have not been recommended in the review literature. In addition, the current thesis presented this issue transparently, and complementary work, the content analysis in Paper 4, which is not subject to these common-method effects was also presented.

Fifth, though there is no clear alternative, the subjectivity inherent to selecting alternative operationalizations for the multiverse analytical approach in Paper 2, and content analysis in Paper 4 is also a limitation to some extent. As has been pointed out for multiverse approaches in general, what are considered valid alternative approaches will vary between researchers (Simonsohn et al., 2020). However, I argue the value of considering these effects and providing analytical examples outweigh potential problems. Both papers also include extensive supplementary material, including for Paper 2, synthetic data, information on how items were selected, and stored results (see https://osf.io/rxv5q/ and https://osf.io/k7qth/). This level of transparency therefore also allows those with, for instance, concerns about the equivalence of alternative indicators to consider findings in light of decisions made. For both Papers 2 and 4, this work considering indicator types and assigning them into domains, was also conducted by all members of the supervisory team, limiting individual subjectivity. Furthermore, for Paper 4 existing approaches to content analysis were also available (Fried, 2017; Newson et al., 2020) and followed.

Findings in Light of a Fast-Paced Methodological Field

In addition to the limitations above, I also present how the papers of the thesis might be interpreted in light of newer work. As noted in Chapter 3, psychometric, particularly network, methods are evolving extremely quickly. This means it is worth drawing attention to developments that have occurred since the publication of the papers which might have affected the approach or interpretation of findings. A thorough review of all recent literature is beyond the scope of this chapter, but two notable papers are highlighted.

First, work by Rhemtulla et al. (2020) demonstrating the potential bias introduced by estimating latent factors in some instances could have implications for Paper 1. The authors demonstrated that in factor models where unique item variance is not purely random error, latent factor correlations can be *inflated*. One of the aims of Paper 1 was to correct for the potential *deflation* of omitting measurement error (i.e. unique variance) in some previous dual-factor studies that relied on observed scores. It is therefore possible some would now argue the relatively high correlations between constructs found in Paper 1 should be treated with some caution. Nevertheless, the results and conclusions of Paper 1 likely remain valid for two reasons: First, the latent correlations reported were in line with previous results at the observed-score level (Antaramian et al., 2010; Keyes, 2005; Patalay & Fitzsimons, 2018; Suldo et al., 2011), though, as expected, they were higher than observed correlations in the same data (see Paper 1, Tables 1 and 4). Second, factor loadings were relatively high. While Rhemtulla et al. (2020), do not provide specific guidance, they emphasize the risk of inflation of factor correlations is particularly high when loadings are low (meaning the error is high).

Rhemtulla et al. (2020) also particularly highlight that this factor correlation inflation has especially problematic consequences for structural models, as bias for structural parameters can be introduced in *both* directions. Their work therefore suggests the approach in Paper 1 may be more risky when included in structural models, and could limit the reported correlations estimated between mental health factors and gender and income (which were not a primary focus of the paper). This work also supports the move to ESEM and network models in subsequent papers in the thesis since these do not carry the same risks.

The second notable addition to the literature is by Neal and Neal (2021). In their recent paper these authors describe an issue that has been known in the social network literature for some time, but which they argue has not been adequately considered in psychological networks. This is the boundary specification problem, in which the exclusion of certain indicators from the universe of a given domain leads to substantial bias in parameters in the estimated network. Many psychological network studies have arguably tackled this by focussing on a given disorder (Robinaugh et al., 2020), by default rooting the domain in a diagnostic psychiatric framework. However, as described in Chapter 1, this can be problematic given issues such as reliability and comorbidity and relies on the unrealistic correspondence theory of truth (Kendler, 2016).

Given the finding in Paper 1 that externalizing symptoms were also strongly related to internalizing symptoms, it is possible the exclusion of externalizing problems in Paper 2 could bias parameters. However, since the universe of general mental health is unclear (see conceptual issues above), the fact internalizing/emotional indicators seem to be theoretically and empirically important (Papers 1 and 4), the prevalence of problems in this area, and apparent sensitivity of adolescence for these (NHS Digital, 2018; Rapee et al., 2019), the approach in Paper 2 makes an important contribution. Interestingly, this issue highlighted by Neal and Neal (2021) also suggests more bottom-up work with key stakeholders would be useful to inform which indicators should be included, mirroring issues highlighted in Chapter 1, and Papers 2-4.

Recommendations and Future Directions

Individual recommendations and future directions were introduced throughout the discussion above. Here I provide a summary of these drawing across findings of the thesis.

Psychometric Researchers

The poor standards of measure development, conceptual confusion and general low psychometric quality found have several implications. First, there is a clear need to provide additional validity evidence, particularly content validity, for available measures and to inform new development. This should consider conceptualization but also prioritize comprehensibility and age-appropriateness. Items such as “I am lonely” or “I worry a lot” which represent theoretically and empirically important indicators *and* are simply worded should be prioritised. Consultation with experts, and particularly young people, is vital to clarify relevant indicators and ensure items are age appropriate and consistently interpreted across groups. The current thesis has identified certain indicators which likely can be uncontroversially included, but also identified others as particular candidates for consideration by young people and experts (e.g., loneliness and self-esteem).

Similarly, the thesis provided initial evidence that emotional/internalizing indicators and constructs may be theoretically and empirically important, but this should be explored further in future work. Second, a more detailed review of content validity evidence for existing measures could further clarify which measures are promising from this perspective. This would build on the relatively brief approach adopted in Paper 4, given the overall scope of this paper, and established frameworks to do this more comprehensively are available (Terwee et al., 2018).

Third, robust single measures that capture both positive and negative indicators are needed so that the relationship of these experiences can be properly evaluated. Such measures would allow better consideration of meaningful covariance, avoiding that potentially introduced by implementation properties such as response format (Podsakoff et al., 2003). Fourth, in order to do this, careful work, drawing on that to improve conceptualization with young people and experts, would also be needed to determine scoring approaches. This would include developing strategies to accommodate effects introduced by reverse wording, and consider theoretical, practical and empirical trade-offs for sumscoring.

Fifth, conceptualization would also be aided by work considering the relative sensitivity ofindicators, providing empirical support at this level to complement qualitative work. This could build on the initial evidence provided here for key indicators and include formal models (Haslbeck et al., 2021). Sixth, where this is done considering within-person longitudinal data, shorter lags than a year between waves could provide useful insight, since Paper 2 found many interactions between indicators to happen more quickly than this, as would be expected for mental health (Epskamp et al., 2018).

Applied Researchers and Practitioners

Seventh, where existing measures are used, outcomes within general mental health should be carefully selected in terms of theory, and judiciously analysed. If several scales are used to provide more comprehensive insight, theoretical and empirical overlap should be explicitly assessed. Eighth, such decisions and evaluation of measures must be transparently reported to aid interpretation and future work. Papers 1-3 provide examples of approaches to this. Ninth, it is likely that positive and negative approaches will both provide useful insight, so given that judicious and transparent assessment of measures takes place, researchers and practitioners could draw on both, depending on their contextual aims and needs.

Tenth, researchers and practitioners should hold measures to higher standards, avoiding heuristics such as coefficient alpha, but also heeding wider issues such as the age-appropriateness problems raised in Paper 3. To do this, reviews drawing on established systems like the COSMIN ratings in Paper 4 should be consulted, rather than reviews which gave more basic information on psychometric properties (e.g., Deighton et al., 2014). Scoring is likely a particular issue given the structural validity and conceptual homogeneity issues uncovered in Paper 4, suggesting that many widely used measures such as the SDQ should not be analysed assuming scores and subdomains are valid. Researchers' initial steps to evaluate measures, mentioned above, should therefore include checks on scoring and structure.

Finally, relevant to all of these recommendations, researchers should try to avoid falling prey to jangle fallacies, e.g. that measures with different names, measure different things (Marsh, 1994). This has arguably been perpetuated in much dual-factor work because of failures to appropriately select, analyse and report constructs and measures. Key issues (Papers 3 and 4) and examples of robust approaches to using available measures (Papers 1 and 2) have been provided in the current thesis as groundwork for future research.

Conclusion

The current thesis provided evidence that poor standards in adolescent general mental health measurement are common and need to be addressed. Conceptual and psychometric problems were highlighted as well as opportunities. Overall, the papers of the thesis suggest considerable work is needed to develop more useful general population measures. However, examples of how to more robustly consider existing measures and datasets were also presented. Positive and negative approaches to adolescent general mental health could be useful for screening and estimation of prevalence but, where problematic measures are used, judicious analysis, following examples and issues set out here, are urgently needed, before accurate inferences can be made. It is likely that studies and new measures should include simple items focussing on *happiness* and *worry*. Several other indicators also showed tentative evidence of being important, but work with young people and experts is needed to explore these and other indicators further. Taken together, the current thesis provides a major step forward for adolescent self-report general mental health measurement, by highlighting ingrained issues, and approaches to interrogate, accommodate and improve these.

References

Aalbers, G., McNally, R. J., Heeren, A., de Wit, S., & Fried, E. I. (2019). Social media and depression symptoms: A network perspective. *J Exp Psychol Gen*, *148*(8), 1454-1462. https://doi.org/10.1037/xge0000528

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5* (Fifth edition. ed.). American Psychiatric Association.

Antaramian, S. P., Huebner, E. S., & Valois, R. F. (2008). Adolescent Life Satisfaction. *Applied Psychology*, *57*(s1), 112-126. https://doi.org/https://doi.org/10.1111/j.1464-0597.2008.00357.x

Antaramian, S. P., Huebner, S. E., Hills, K. J., & Valois, R. F. (2010). A Dual-Factor Model of Mental Health: Toward a More Comprehensive Understanding of Youth Functioning. *American Journal of Orthopsychiatry*, *80*(4), 462-472. https://doi.org/10.1111/j.1939-0025.2010.01049.x

Bartels, M., Cacioppo, J. T., van Beijsterveldt, T. C. E. M., & Boomsma, D. I. (2013). Exploring the Association Between Well-Being and Psychopathology in Adolescents. *Behavior Genetics*, *43*(3), 177-190. https://doi.org/10.1007/s10519-013-9589-7

Black, L., Panayiotou, M., & Humphrey, N. (2020). The special relationship of internalizing symptoms and wellbeing: A cross-validation study considering indicator-level associations beyond the dual-factor model of mental health. https://doi.org/https://doi.org/10.31234/osf.io/stajk

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440. https://doi.org/10.1007/s11336-006-1447-6

Carlson, K. D., & Herdman, A. O. (2012). Understanding the Impact of Convergent Validity on Research Results. *Organizational Research Methods*, *15*(1), 17-32. https://doi.org/10.1177/1094428110392383

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, *25*(3), 259-270. https://doi.org/10.1037/met0000236

Cole, A., Bond, C., Qualter, P., & Maes, M. (2021). A Systematic Review of the Development and Psychometric Properties of Loneliness Measures for Children and Adolescents. *International* *journal of environmental research and public health*, *18*(6), 3285. https://www.mdpi.com/1660-4601/18/6/3285

Costello, J. (2015). Commentary: ‘Diseases of the world’: from epidemiology to etiology of child and adolescent psychopathology – a commentary on Polanczyk et al. (2015). *Journal of Child Psychology and Psychiatry*, *56*(3), 366-369. https://doi.org/10.1111/jcpp.12402

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The Validity of the Multi-Informant Approach to Assessing Child and Adolescent Mental Health. *Psychological Bulletin*, *141*(4), 858-900. https://doi.org/10.1037/a0038498

Deighton, J., Croudace, T., Fonagy, P., Brown, J., Patalay, P., & Wolpert, M. (2014). Measuring mental health and wellbeing outcomes for children and adolescents to inform practice and policy: a review of child self-report measures. *Child and Adolescent Psychiatry and Mental Health*, *8*(1), 14. https://doi.org/10.1186/1753-2000-8-14

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259-268. https://doi.org/https://doi.org/10.1016/j.jrp.2011.02.004

Epskamp, S. (2019). Reproducibility and Replicability in a Fast-Paced Methodological World. *Advances in Methods and Practices in Psychological Science*, *2*(2), 145-155. https://doi.org/10.1177/2515245919847421

Epskamp, S., van Borkulo, C. D., van der Veen, D. C., Servaas, M. N., Isvoranu, A.-M., Riese, H., & Cramer, A. O. J. (2018). Personalized Network Modeling in Psychopathology: The Importance of Contemporaneous and Temporal Connections. *Clinical Psychological Science*, *6*(3), 416-427. https://doi.org/10.1177/2167702617744325

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456-465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research:Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378. https://doi.org/10.1177/1948550617693063

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191-197. https://doi.org/https://doi.org/10.1016/j.jad.2016.10.019

Fried, E. I. (2020). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, *31*(4), 271-288. https://doi.org/10.1080/1047840X.2020.1853461

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1), 72. https://doi.org/10.1186/s12916-015-0325-4

Fried, E. I., Nesse, R. M., Zivin, K., Guille, C., & Sen, S. (2014). Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. *Psychological Medicine*, *44*(10), 2067-2076. https://doi.org/10.1017/S0033291713002900

Fritz, J., de Graaff, A. M., Caisley, H., van Harmelen, A.-L., & Wilkinson, P. O. (2018). A Systematic Review of Amenable Resilience Factors That Moderate and/or Mediate the Relationship Between Childhood Adversity and Mental Health in Young People. *Frontiers in Psychiatry*, *9*(230). https://doi.org/10.3389/fpsyt.2018.00230

Furlong, M. J., Fullchange, A., & Dowdy, E. (2017). Effects of mischievous responding on universal mental health screening: I love rum raisin ice cream, really I do! *School psychology quarterly : the official journal of the Division of School Psychology, American Psychological Association*, *32*(3), 320-335. https://doi.org/10.1037/spq0000168

Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence*, *62*, 138-147. https://doi.org/https://doi.org/10.1016/j.intell.2017.04.001

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to Use Broader Internalising and Externalising Subscales Instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British Parents, Teachers and Children. *Journal of Abnormal Child Psychology*, *38*(8), 1179-1191. https://doi.org/10.1007/s10802-010-9434-x

Graber, J. A., & Sontag, L. M. (2009). Internalizing Problems During Adolescence. In *Handbook of Adolescent Psychology*. https://doi.org/https://doi.org/10.1002/9780470479193.adlpsy001020

Greenspoon, P. J., & Saklofske, D. H. (2001). Toward an Integration of Subjective Well-Being and Psychopathology. *Social Indicators Research*, *54*(1), 81-108. https://doi.org/10.1023/a:1007219227883

Haeffel, G. J., Jeronimus, B. F., Kaiser, B. N., Weaver, L. J., Soyster, P. D., Fisher, A. J., . . . Lu, W. (2021). Folk Classification and Factor Rotations: Whales, Sharks, and the Problems With the Hierarchical Taxonomy of Psychopathology (HiTOP). *Clinical Psychological Science*, *0*(0), 21677026211002500. https://doi.org/10.1177/21677026211002500

Hallquist, M. N., Wright, A. G. C., & Molenaar, P. C. M. (2019). Problems with Centrality Measures in Psychopathology Symptom Networks: Why Network Psychometrics Cannot Escape Psychometric Theory. *Multivariate Behavioral Research*, 1-25. https://doi.org/10.1080/00273171.2019.1640103

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2021). Modeling psychopathology: From data models to formal theories. *Psychological Methods*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000303

Humphrey, N., & Wigelsworth, M. (2016). Making the case for universal school-based mental health screening. *Emotional and Behavioural Difficulties*, *21*(1), 22-42. https://doi.org/10.1080/13632752.2015.1120051

Iasiello, M., & Agteren, J. v. (2020). *Mental health and/or mental illness: A scoping review of the evidence and implications of the dual-continua model of mental health*. Exeley. https://doi.org/10.3316/informit.261420605378998

Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry*, *15*(1), 5-12. https://doi.org/10.1002/wps.20292

Keyes, C. L. M. (2005). Mental Illness and/or Mental Health? Investigating Axioms of the Complete State Model of Health. *Journal of Consulting and Clinical Psychology*, *73*(3), 539-548. https://doi.org/10.1037/0022-006X.73.3.539

Kline, R. (2015). Principles and practice of structural equation modeling Fourth Edition. In: New York: The Guilford Press.

Krause, K. R., Chung, S., Sousa Fialho, M. d. L., Szatmari, P., & Wolpert, M. (2021). The challenge of ensuring affordability, sustainability, consistency, and adaptability in the common metrics agenda. *The Lancet Psychiatry*, *8*(12), 1094-1102. https://doi.org/https://doi.org/10.1016/S2215-0366(21)00122-X

Marsh, H. W. (1994). Sport Motivation Orientations: Beware of Jingle-Jangle Fallacies. *Journal of Sport and Exercise Psychology*, *16*(4), 365-380. https://doi.org/10.1123/jsep.16.4.365

McElroy, E., Shevlin, M., & Murphy, J. (2017). Internalizing and externalizing disorders in childhood and adolescence: A latent transition analysis using ALSPAC data. *Comprehensive Psychiatry*, *75*, 75-84. https://doi.org/https://doi.org/10.1016/j.comppsych.2017.03.003

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412-433. https://doi.org/10.1037/met0000144

Millsap, R. E. (2012). *Statistical Approaches to Measurement Invariance*. Routledge.

Moeller, J. (2022). Averting the next credibility crisis in psychological science: Within-person methods for personalized diagnostics and intervention. *Journal for Person-Oriented Research*, *7*(2), 53-77. https://doi.org/10.17505/jpor.2021.23795

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonson, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2018). *COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) user manual Version 1.0*. COSMIN. https://cosmin.nl/wpcontent/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf

Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*(4), 201-218. https://doi.org/10.1207/s15366359mea0204_1

Moore, S. A., Dowdy, E., Nylund-Gibson, K., & Furlong, M. J. (2019). A latent transition analysis of the longitudinal stability of dual-factor mental health in adolescence. *Journal of School Psychology*, *73*, 56-73. https://doi.org/https://doi.org/10.1016/j.jsp.2019.03.003

Neal, Z. P., & Neal, J. W. (2021). Out of bounds? The boundary specification problem for centrality in psychological networks. *Psychological Methods*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000426

Newson, J. J., Hunter, D., & Thiagarajan, T. C. (2020). The Heterogeneity of Mental Health Assessment. *Frontiers in Psychiatry*, *11*(76). https://doi.org/10.3389/fpsyt.2020.00076

NHS Digital. (2018). *Mental Health of Children and Young People in England, 2017 Summary of key findings*. https://files.digital.nhs.uk/F6/A5706C/MHCYP%202017%20Summary.pdf

Patalay, P., & Fitzsimons, E. (2016). Correlates of Mental Illness and Wellbeing in Children: Are They the Same? Results From the UK Millennium Cohort Study. *J Am Acad Child Adolesc Psychiatry*, *55*(9), 771-783. https://doi.org/10.1016/j.jaac.2016.05.019

Patalay, P., & Fitzsimons, E. (2018). Development and predictors of mental ill-health and wellbeing from childhood to adolescence. *Social Psychiatry and Psychiatric Epidemiology*, *53*, 1311–1323 https://doi.org/10.1007/s00127-018-1604-0

Patalay, P., & Fried, E. I. (2020). Editorial Perspective: Prescribing measures: unintended negative consequences of mandating standardized mental health measurement. *Journal of Child Psychology and Psychiatry*, *62*(8). https://doi.org/10.1111/jcpp.13333

Percy, A., McCrystal, P., & Higgins, K. (2008). Confirmatory Factor Analysis of the Adolescent Self-Report Strengths and Difficulties Questionnaire. *24*(1), 43-48. https://doi.org/10.1027/1015-5759.24.1.43

Perlman, D., & Peplau, L. A. (1984). Loneliness research: A survey of empirical findings. *Preventing the harmful consequences of severe and persistent loneliness*, *13*, 46.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879-903. https://doi.org/10.1037/0021-9010.88.5.879

Putwain, D. W., Stockinger, K., von der Embse, N. P., Suldo, S. M., & Daumiller, M. (2021). Test anxiety, anxiety disorders, and school-related wellbeing: Manifestations of the same or different constructs? *Journal of School Psychology*, *88*, 47-67. https://doi.org/https://doi.org/10.1016/j.jsp.2021.08.001

Qualter, P., Vanhalst, J., Harris, R., Van Roekel, E., Lodder, G., Bangee, M., . . . Verhagen, M. (2015). Loneliness Across the Life Span. *Perspectives on Psychological Science*, *10*(2), 250-264. https://doi.org/10.1177/1745691615568999

Rapee, R. M., Oar, E. L., Johnco, C. J., Forbes, M. K., Fardouly, J., Magson, N. R., & Richardson, C. E. (2019). Adolescent development and risk for the onset of social-emotional disorders: A review and conceptual model. *Behaviour Research and Therapy*, *123*, 103501. https://doi.org/https://doi.org/10.1016/j.brat.2019.103501

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. *American Journal of Psychiatry*, *170*(1), 59-70. https://doi.org/10.1176/appi.ajp.2012.12070999

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor Models and Rotations: Exploring the Extent to Which Multidimensional Data Yield Univocal Scale Scores. *Journal of Personality Assessment*, *92*(6), 544-559. https://doi.org/10.1080/00223891.2010.496477

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30-45. https://doi.org/10.1037/met0000220

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R., & Borsboom, D. (2020). The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*, *50*(3), 353-366. https://doi.org/10.1017/S0033291719003404

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208-1214. https://doi.org/10.1038/s41562-020-0912-z

Solmi, M., Radua, J., Olivola, M., Croce, E., Soardo, L., Salazar de Pablo, G., . . . Fusar-Poli, P. (2021). Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-021-01161-7

St Clair, M. C., Neufeld, S., Jones, P. B., Fonagy, P., Bullmore, E. T., Dolan, R. J., . . . Goodyer, I. M. (2017). Characterising the latent structure and organisation of self-reported thoughts, feelings and behaviours in adolescents and young adults. *PLOS ONE*, *12*(4), e0175381. https://doi.org/10.1371/journal.pone.0175381

Suldo, S., & Shaffer, E. J. (2008). Looking Beyond Psychopathology: The Dual-Factor Model of Mental Health in Youth. *School Psychology Review*, *37*(1), 52-68.

Suldo, S., Thalji, A., & Ferron, J. (2011). Longitudinal academic outcomes predicted by early adolescents' subjective well-being, psychopathology, and mental health status yielded from a dual factor model. *The Journal of Positive Psychology*, *6*(1), 17-30. https://doi.org/10.1080/17439760.2010.536774

Sydney, I., & Pyle, E. (2018). *Cognitive testing of loneliness questions and response options*. https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/compendium/nationalmeasurementofloneliness/2018/cognitivetestingoflonelinessquestionsandresponseoptions#overallrecommendations

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., . . . de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34-42. https://doi.org/https://doi.org/10.1016/j.jclinepi.2006.03.012

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., . . . Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patientreported outcome measures: a Delphi study. *Quality of Life Research*, *27*(5), 1159-1170. https://doi.org/10.1007/s11136-018-1829-0

van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. J. (2017). What is the p-factor of psychopathology? Some risks of general factor modeling. *Theory & Psychology*, *27*(6), 759-773. https://doi.org/10.1177/0959354317737185

Vostanis, P. (2006). Strengths and Difficulties Questionnaire: Research and clinical applications. *Current Opinion in Psychiatry*, *19*(4), 367-372. https://doi.org/10.1097/01.yco.0000228755.72366.05

Weijters, B., & Baumgartner, H. (2012). Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research*, *49*(5), 737-747. https://doi.org/10.1509/jmr.11.0368

Wolpert, M., Cheng, H., & Deighton, J. (2015). Measurement Issues: Review of four patient reported outcome measures: SDQ, RCADS, C/ORS and GBO – their strengths and limitations for clinical use and service evaluation. *20*(1), 63-70. https://doi.org/doi:10.1111/camh.12065

Yarkoni, T. (2020). Implicit Realism Impedes Progress in Psychology: Comment on Fried (2020). *Psychological Inquiry*, *31*(4), 326-333. https://doi.org/10.1080/1047840X.2020.1853478

Appendix 1: Cross-Sectional Network Paper

The special relationship of internalizing symptoms and wellbeing: A cross-validation study considering indicator-level associations beyond the dual-factor model of mental health

Louise Black, Margarita Panayiotou, and Neil Humphrey

University of Manchester

Author Note

Louise Black, Manchester Institute of Education, University of Manchester; Margarita Panayiotou, Manchester Institute of Education, University of Manchester;
Neil Humphrey, Manchester Institute of Education, University of Manchester.

This research was supported by grants from The National Lottery Community Fund, using data from the HeadStart Evaluation.

Correspondence concerning this article should be addressed to Louise Black,

Manchester Institute of Education, University of Manchester, Manchester, M13 9PL. E-mail: louise.black@manchester.ac.uk

Abstract

Dual-factor models of mental health have emphasized the importance of considering both mental health difficulties and wellbeing in young people. However, studies to date have failed to consider the conceptual and statistical similarity between wellbeing and internalizing symptom indicators. Drawing on latent variable and network methods, we present exploratory and confirmatory models of indicator-level interactions for internalizing symptoms and wellbeing in two large independent samples (sample 1, *N* = 14,805, sample 2, *N* = 14,066). Exploratory and confirmatory networks fitted notably better than the factor model. A densely connected network was found which showed good stability across samples. The most and strongest relationships were observed *within* each domain. Nevertheless, direct associations similar to some within-domain relationships were also observed *between* internalizing and wellbeing indicators. Worry was the most central indicator in both samples, and was meaningfully related to the wellbeing indicator “feeling relaxed”. Evidence of known-groups validity was demonstrated since indicator-level interactions in the networks were sensitive to known differences between girls and boys. Findings suggest the covariance between internalizing symptom and wellbeing indicators in adolescent datasets should be more closely attended to, and controlled for in intervention and epidemiological research.

Key words: internalizing symptoms, wellbeing, adolescence, network analysis

The current study

Since evidence of the dual-factor model to date has tended to rely on problematic inference (Moore et al., 2019), we aimed to shed light on psychometric properties underpinning the theory. To situate the particular measures used in the wider literature, we first aimed to replicate the substantial negative covariance found between internalizing symptoms and wellbeing at the latent variable level. Moving beyond the constraints of latent variable models (LVMs), we also hypothesized there would be substantial indicator-level covariance within and between domains in line with network theory (Borsboom, 2017) and in light of the conceptual similarity of the constructs (Alexandrova & Haybron, 2016). We also predicted indicators in the network would cluster into two domains, reflecting the measures they were drawn from. In order to gain further insight into whether particular indicators within the network were particularly important, we also explored strength centrality (Bringmann et al., 2019). Finally, given the reported gender differences in internalizing symptoms, known-groups validity of the network was considered. We hypothesized being a girl would be positively associated with symptom and negatively associated with wellbeing indicators.

Method

Background and Procedure

Secondary analysis was conducted of a large community study in England that aims to explore and test new ways to improve mental health and wellbeing of young people aged 10–16 and prevent serious mental health issues from developing. Data are collected on an annual basis from pupils in year nine (grade 8; age 13-14). Ethical approval was granted (reference masked for review), and opt-out parental consent was given for adolescents to complete secure online surveys at school. Teachers read out an information sheet which emphasized pupils' confidentiality and right to withdraw.

Participants

We used the two year nine samples from the first two time points (sample 1 [S1], *N* = 14,805, sample 2 [S2], *N* = 14,066). These sample sizes reflect the number of participants from the main dataset who did not have missing data on all items analysed here (for this reason 177 were excluded from S1 and 175 from S2). Data were collected in 2017 and 2018 for S1 and S2, respectively. Schools involved in the project were the same at each time point, with pupils sampled from 112 schools for S1 and 105 for S2. Characteristics of each sample are provided in Table 1 with reference to national figures and published norms (Department for Education, 2017a, 2017b, 2017c, 2018a, 2018b, 2018c; Ng Fat et al., 2017; Youth in Mind, 2016).

Measures

Internalizing symptoms were assessed by the five-item self-report emotional symptoms scale from the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997), on a three-point scale (*not true, somewhat true, certainly true*). Higher scores indicate more emotional difficulties. This subscale has acceptable reliability and known-groups validity in 11–16 year-olds (Goodman et al., 1998). Wellbeing was measured by the 7-item Short Warwick-Edinburgh Mental Wellbeing Scale (SWEMWBS; Stewart-Brown et al., 2009), which has a five-point scale (*none of the time, rarely, some of the time, often, all of the time*). Higher scores represent higher wellbeing. This measure has shown good fit, and gender and year-group invariance in 13–16 year-olds (Hunter et al., 2015). In the current sample internal consistency was good for emotional symptoms with Cronbach's α = .75 and ordinal alpha based on the polychoric matrix (Gadermann et al., 2012) α = .82 in both samples. For SWEMWBS Cronbach's was α = .84 and ordinal alpha was α = .86 in both samples.

Abbreviated item wording can be seen in Table 2.

Statistical Analysis

Analyses were conducted in R 3.6.1 (code is provided in the supplementary material).

Confirmatory factor analysis.

A correlated LVM was estimated via confirmatory factor analysis in S1. The five SDQ items loaded onto an emotional symptoms factor while the seven SWEMWBS items loaded onto a wellbeing factor. The LVM was estimated in the *lavaan* package (version 0.65; Rosseel et al., 2020) using the weighted least squares with mean and variance adjusted estimator to account for the ordinal nature of the data. Chi-square statistics are reported but not interpreted as indicative of fit given their sensitivity to sample size. Fit was otherwise interpreted according to recommended thresholds with good fit considered for the comparative fit index (CFI) and Tucker-Lewis index (TLI) above .95, root mean square error of approximation (RMSEA; including 90% confidence intervals, CIs) below .06, and standardized root mean square residual below .08 (Hu & Bentler, 1999).

Exploratory network analysis.

A gaussian graphical model (GGM; Epskamp et al., 2018) was freely estimated for S1 in the *psychonetrics* package (version 0.7.1; Epskamp, 2020). GGMs are undirected networks of partial correlations. Indicators are referred to as nodes, depicted by circles, and partial correlations between these are referred to as edges, depicted by lines. This model allowed consideration of the nature and extent of direct relationships between indicators within and between domains. Weighted least squares estimation was used, treating data as ordinal (i.e. a polychoric matrix and thresholds were estimated), with pairwise deletion. Non-significant parameters were then removed recursively at α = .01 from the resulting saturated model. Model fit was expected to be good (according to the above criteria) given the data-driven approach.

Confirmatory network analysis.

The framework for confirmatory network modeling set out by Kan et al. (2019) was used. Using the pattern of parameters derived in the exploratory model as input, a confirmatory network was estimated in S2. Edges retained in the exploratory model were freely estimated, while the rest were fixed at zero in the confirmatory model (via the adjacency matrix). Descriptive statistics and average differences between retained edges were calculated for the two networks.

Network characteristics.

Given, the conditional nature of relationships and typical density of psychological data (Williams & Rast, 2019), edge weights were not expected to be large, and considered in line with Cohen's (1992) effect sizes for partial correlations (*r* = .02–.15, .15–.35, > .35 as small, medium, large). Clustering of nodes into domains was assessed via the walk-trap algorithm in the *igraph* package (version 1.2.5; Csárdi, 2020) which has been shown to identify groups equivalent to factors and is more accurate than traditional methods such as parallel analysis (Golino & Epskamp, 2017). Strength centrality, the sum of edge weights related to the target node (Costantini et al., 2015), was also computed for each network. Nodes with higher strength are those most influenced by others, or the most influential in the network. Strength is also highly correlated with the amount of shared variance between nodes

(Costantini et al., 2015).

Known-groups validity testing.

Following Christensen et al. (2020), the known group covariate gender was added to the S1 model, and replicated using the S1 adjacency matrix in S2. This allowed consideration of whether girls and boys differed for particular indicators and therefore provided insight into the possible utility of considering indicator-level covariance.

Results

Missingness was low in S1 (1.3-3.8%) and S2 (1.1-4.5%). Descriptive statistics are shown in Table 2. The LVM in S1 showed mixed fit, $\chi^2(53) = 3724.66$, $p < .001$, CFI = .964, TLI = .955, RMSEA = .073 (90% CI [071, .075]), SRMR = .044, with a strong negative correlation between the two constructs (r = -.59; the other parameters are shown in the supplementary material Figure S1).

As expected, the exploratory network model fitted very well, $\chi^2(18) = 36.29$, $p = .006$, CFI > .999, TLI = .998, RMSEA = .008 (90% CI [.004, .012]). Fit remained good for the confirmatory model, $\chi^2(18) = 116.14$, $p < .001$, CFI = .997, TLI = .991, RMSEA = .020 (90% CI [.016, .023]), though three edges in the network were no longer significant at .01. These edges were between nervous and somatic, somatic and useful, and scared and optimism (ranging from r = .03-.05 in S1). These could therefore have been retained due to sampling variability in S1. Exploratory and confirmatory networks are also plotted in Figure 1 A and C, with thicker edges representing stronger relationships (scaled across the two networks) and blue/red lines indicating positive/negative partial correlations. Full matrices of edge weights for both samples are provided in the supplementary material in Tables S1 and S2.

The mean difference between equivalent edges in the two networks was .01 (*SD* = .01). As expected, the walk-trap algorithm identified two communities which corresponded exactly to the two scales used. Descriptive statistics of the two networks (see Table 3) confirmed that within domain portions of the network were denser and showed stronger relationships. However, it can also be seen from Figure 1 and the supplementary material that some between-domain edges were of equivalent magnitude to those within domains. The rank order and magnitude of strength for each node was similar for both networks, with worry consistently the most central (see Figure 2). Worry also showed the largest cross-domain relationship (exploratory, r = -.18, confirmatory, r = -.17).

When gender was added to the S1 network (see Figure 1 B) to consider known-groups validity, $\chi^2(25) = 38.77$, $p = .04$, CFI > .999, TLI = .999, RMSEA = .006 (90% CI [.001, .010]), small positive associations (r = .05-.14) were found between being a girl and the indicators close to people, somatic, worry, nervous and scared, and a negative edge was found between being a girl and relaxed (r = -.13). In terms of network density, a similar level was seen once gender was controlled for, with 80.30% of possible parameters retained. The S2 gender validity network (Figure 1 D) fitted well, $\chi^2(25) = 139.40$, $p < .001$, CFI > .998, TLI = .991, RMSEA = .018 (90% CI [.015, .021]), though the same three edges as in the model without gender were not significant at .01. The estimated edges between gender and indicators mentioned above also remained similar (r = .04-.16). Full matrices of edge weights for both samples including gender are provided in the supplementary material in Tables S3 and S4.

Discussion

While dual-factor theories of mental health seem to have informed fields such as epidemiological research (e.g., NHS Digital, 2018) and social and emotional learning (Humphrey, 2013), little consideration has been given to the joint measurement of symptoms and wellbeing. Specifically, we argue the conceptual and statistical similarity between internalizing symptoms and wellbeing should not be ignored in analysis. The current study confirmed the substantial correlation between these constructs at the latent variable level, before considering exploratory and confirmatory networks between indicators in large independent samples. The network appeared to fit substantially better than the LVM, suggesting the constraints on indicator covariance in the latter were likely unreasonable. This is consistent with theory that internalizing problems such as depression may be particularly likely to result from symptoms causing one another rather than latent disease processes (Fried & Cramer, 2017). It is also in line with the idea that LVMs are likely too strict in general for psychological data (Cramer et al., 2012). The network also showed mostly good stability across samples, with excellent fit in the confirmatory model though a small number of parameters were non-significant. Though indicators clustered by measure, considerable meaningful covariance was found between domains. Evidence of known-groups validity was suggested by meaningful associations with gender, and we demonstrated for the first time that indicator-level consideration may be necessary to accurately capture gender differences. The correlation between internalizing and wellbeing at the latent construct level replicates previous work (Black et al., 2019). This relationship is therefore consistent across different age groups (10-11 year-olds versus 13-14 year-olds) and measures considered thus far, including different operationalizations of wellbeing (life satisfaction and eudaimonia). A correlation of this magnitude suggests a possible common cause (Reise et al., 2010), as was modelled elsewhere (Black, 2019). While common cause models can be useful to consider the degree of dimensionality (Rodriguez et al., 2016), they do not easily provide information about indicator-level covariance.[9]

This was the first study, to our knowledge, to model indicator-level associations. We found a densely connected network, including between internalizing symptoms and wellbeing, which are often considered to be distinct. This indicates item responses to these domains should not be considered to be independent. In addition, no large relationships were found, even within domains, suggesting each item captured something distinct, as would be expected given the conceptual content. It may therefore be questionable to consider each item as interchangeable indicators of the same construct, as total scores do (McNeish & Wolf, 2020).

The LVM also indicated that when latent variables were assumed, some indicators were influenced up to 1.5 times as much by the latent trait as others. Consistent with simulation evidence demonstrating the equivalence of network strength and factor loadings (Hallquist et al., 2019), the indicators with highest strength, worry (emotional symptoms) and thinking clearly (wellbeing), also had high factor loadings, suggesting these relate strongly to others and may therefore be particularly useful indicators. The medium effect between relaxed and worry may also suggest these are particular risk factors for one another. However, it may also be that they capture similar states and in fact difficulty relaxing is used as an indicator of generalized anxiety disorder (Spitzer et al., 2006). Unhappy was also directly related to all but one of the wellbeing indicators, demonstrating a particular link with affective symptoms and wellbeing (in this case eudaimonic). If a more hedonic operationalization of wellbeing had been considered (i.e. direct inclusion of happiness via affect or life satisfaction), these associations may have been even stronger.

This perhaps speaks to an issue in the broader mental health measurement field, that there is often overlap of indicators between, and heterogeneity within domains (Newson et al., 2020). This means the choice of measure likely causes substantial variability in results (Yarkoni, 2020). Considering the state of the field as a whole (Newson et al., 2020), and the varied operationalizations

[9] While correlations between items’ unique variances can be accommodated in factor models, this tends to only be possible via methods tantamount to *p*-hacking (Pan et al., 2017) .

of each domain, it is possible differences between internalizing symptom and wellbeing indicators are sometimes similar to differences within either domain. This was supported in the current study, with the relationship between feeling relaxed and worry at a similar magnitude to relationships within both the internalizing symptom and wellbeing domains. This issue should therefore be considered in future work.

It also remains unclear what valid variance is common or not between the constructs. While dual-factor research has tended to treat these outcomes as distinct, elsewhere they have been considered interchangeable (e.g., Orben & Przybylski, 2019).We urge researchers and policy makers to consider measurement issues such as those highlighted in the present paper more prominently when designing studies and screening programs (see conclusion and recommendations below).

Despite this lack of clarity in conceptualization and measurement, some work has examined whether distinct correlates were associated with symptoms and wellbeing, or different groups of complete mental health (e.g., Grych et al., 2020; Patalay & Fitzsimons, 2016; Suldo et al., 2016). For instance, cognitive ability was found to be associated with symptoms but not wellbeing, while obesity was associated with only wellbeing (Patalay & Fitzsimons, 2016). Such approaches are therefore appealing to understand how to target intervention. However, based on the evidence reviewed here and our findings, more work is needed to understand whether each construct validly measures something distinct before considering the dissociation of correlates based on analyses where symptoms and wellbeing are forced to be separate.

We provided preliminary insight into this issue by considering known-groups validity for girls and boys (Merikangas et al., 2010; NHS Digital, 2018). Our analysis revealed certain (but not all) symptoms were directly related to being a girl. Considering indicator-level interactions can therefore provide further novel insight into the known differences between girls and boys. The wellbeing indicator, feeling relaxed, also showed a negative association with being a girl, suggesting this item captured similar risk to certain internalizing items. However, the wellbeing indicator feeling close to others, was *positively* associated with being a girl. This is consistent with findings elsewhere, that girls tend to report higher perceived social support (Rueger et al., 2010). Similarly, small differences have been found between girls and boys for traits such as agreeableness and warmth (Perry & Pauletti, 2011). Generally higher self-reported levels of such personality factors in girls may therefore affect responding and positive association found here. It has also been suggested such interpersonal strengths can in fact be risk factors for internalizing symptoms in girls, for instance via increased susceptibility to interpersonal stress (Kuehner, 2017). While interpersonal relationships are considered to be part of eudaimonic wellbeing (Tennant et al., 2007), our results suggest total wellbeing scores may mask key complexities. The variation in direction of association with gender may also explain why some studies have found no gender differences for wellbeing (e.g., Patalay & Fitzsimons, 2016), while others have (NHS Digital, 2018): It is possible only some items have salient gender differences, such that those included in a particular scale do not capture a difference, or that these differences at the item level cancel each other out when sum scored.

In order to address the issues discussed thus far, future work should control for common method variance among items of the same scale (Hallquist et al., 2019). In fact, all dual-factor literature, to our knowledge, relies on separate scales to measure symptoms and wellbeing. Since varying response formats can substantially affect results (Weijters et al., 2010), it will be important to explicitly model variance in the item covariance structure attributable to being from the same or different scales. We addressed problems associated with using different scales as far as possible within the network modeling framework, by using a polychoric matrix which considers the underlying continuous distribution of categorical scales. While residual network modeling can theoretically be used to control for common variance before estimating a network (Epskamp et al., 2017), future work will rely on the development of methods to handle ordinal data within this framework, and little is known about whether the method overfits data. Scales are also being developed to tap both positive and negative aspects of mental health together (Alexander et al., 2020; Dowdy et al., 2011; Renshaw & Bolognino, 2017). Using single measures, such as these, might also shed more light on dual-factor research.

Strengths and Limitations.

While many studies have considered symptoms and wellbeing together in young people, this study was the first to explore indicator-level associations between similar internalizing symptom and wellbeing constructs. We considered three core aspects of validity which are often ignored: conceptual, structural and known-groups validity (Flake et al., 2017). This was achieved via cross-validation across two large samples, using robust methods.

Though our findings could be specific to the particular measures used, these are considered to be useful brief measures and are widely used in cohort studies (Johnston & Gowers, 2005; NHS Digital, 2018). Additionally, the correlation at the latent level observed in the current study was similar to elsewhere using different measures (Black et al., 2019). While findings therefore need to be replicated considering other samples (e.g. different age groups) and measures, our findings can be considered to have utility for researchers and policy makers.

Robust methods were used in the current study to handle ordinal data, and significance pruning was used, rather than typically-used lasso estimation, which can result in high false positives for dense networks at large sample sizes (Williams & Rast, 2019). However, the density of the network returned in the exploratory model meant the confirmatory model freely estimated a substantial proportion of the parameters. Nevertheless, parameters in each network appeared to be similar.

The sample was not drawn to be representative, meaning results should only be generalized to similar populations. While the self-report SDQ measure as a whole has also been found to have items that can be difficult to interpret, the subscale used in the current analysis (emotional problems) was not likely to represent major problems for the age-range considered here (Black et al., 2020). Finally, though gender was used as a covariate, measurement invariance across boys and girls was not considered since this was beyond the scope of the current study. However, there is evidence of gender invariance for SWEMWBS (Hunter et al., 2015) and other versions the SDQ (He et al., 2013; Ortuño-Sierra et al., 2015).

Conclusions and Recommendations

We found evidence of considerable associations between indicators of internalizing symptoms and wellbeing via a network in young people. Latent variable-level correlations, accounting for measurement error, also suggested considerable common variance across internalizing and wellbeing though the LVM fitted substantially worse than the confirmatory network model. These findings do not suggest the same construct was captured by the scales used in the current study. Rather, there are likely similarities or risk associations between the indicators. We also found support for the validity of considering indicator-level responses for internalizing and wellbeing since this seemed sensitive to known gender differences and captured nuances consistent with the broader literature. However, we cannot know to what extent method effects create artificial divisions between items of different scales. We therefore suggest covariance should be accounted for in analysis, for instance via correlated factors in structural equation models or via network models. Where simpler modeling strategies are sought, a more conservative approach might be to measure only one outcome, in order to avoid introducing bias into results via unmodeled overlap. Where symptoms are of greater interest, for instance in clinical populations, internalizing scales might be chosen, and when screening or intervention outcomes are considered in the general population, wellbeing might provide greater variability.

References

Alexander, L. M., Salum, G. A., Swanson, J. M., & Milham, M. P. (2020). Measuring strengths and weaknesses in dimensional psychiatry. *Journal of Child Psychology and Psychiatry*, *61*(1), 40-50. https://doi.org/10.1111/jcpp.13104

Alexandrova, A., & Haybron, D. M. (2016). Is Construct Validation Valid? *Philosophy of Science*, *83*(5), 1098-1109. https://doi.org/10.1086/687941

Black, L., Mansfield, R., & Panayiotou, M. (2020). Age Appropriateness of the Self-Report Strengths and Difficulties Questionnaire. *Assessment*, *0*(0), 1073191120903382. https://doi.org/10.1177/1073191120903382

Black, L., Panayiotou, M., & Humphrey, N. (2019). The dimensionality and latent structure of mental health difficulties and wellbeing in early adolescence. *PLOS ONE*, *14*(2), e0213018. https://doi.org/10.1371/journal.pone.0213018

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5-13. https://doi.org/doi:10.1002/wps.20375

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., . . . Snippe, E. (2019). What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology*, *128*(8), 892-903. https://doi.org/10.1037/abn0000446

Christensen, A. P., Golino, H., & Silvia, P. J. (2020). A Psychometric Network Perspective on the Validity and Validation of Personality Trait Questionnaires. *European Journal of Personality*, *34*(6). https://doi.org/10.1002/per.2265

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155. Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13-29. https://doi.org/https://doi.org/10.1016/j.jrp.2014.07.003

Cramer, A. O. J., Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., . . . Borsboom, D. (2012). Measurable Like Temperature or Mereological Like Flocking? On the Nature of

Personality Traits. *European Journal of Personality*, *26*(4), 451-459. https://doi.org/doi:10.1002/per.1879

Csárdi, G. (2020). Package 'igraph'. 1-473. https://cran.r-project.org/web/packages/igraph/igraph.pdf

Department for Education. (2017a). *Pupil premium: allocations and conditions of grant 2016 to 2017*. Retrieved 20/09/2018 from https://www.gov.uk/government/publications/pupil-premium-conditions-of-grant-2016-to-2017

Department for Education. (2017b). *Schools, pupils and their characteristics: January 2017*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650547/SFR28_2017_Main_Text.pdf

Department for Education. (2017c). *Special educational needs in England: January 2017*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633031/SFR37_2017_Main_Text.pdf

Department for Education. (2018a). *Pupil premium: allocations and conditions of grant 2017 to 2018*. https://www.gov.uk/government/publications/pupil-premium-conditions-of-grant-2017-to-2018

Department for Education. (2018b). *Schools, pupils and their characteristics: January 2018*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/719226/Schools_Pupils_and_their_Characteristics_2018_Main_Text.pdf

Department for Education. (2018c). *Special educational needs in England: January 2018*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/729208/SEN_2018_Text.pdf

Dowdy, E., Twyford, J. M., Chin, J. K., DiStefano, C. A., Kamphaus, R. W., & Mays, K. L. (2011). Factor structure of the BASC–2 Behavioral and Emotional Screening System Student Form. *Psychological assessment*, *23*(2), 379-387. https://doi.org/10.1037/a0021843

Epskamp, S. (2020). *Package 'psychonetrics'*. https://cran.r-project.org/web/packages/psychonetrics/psychonetrics.pdf

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized Network Psychometrics: Combining Network and Latent Variable Models. *Psychometrika*, *82*(4), 904-927. https://doi.org/10.1007/s11336-017-9557-x

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian Graphical Model in Cross-Sectional and Time-Series Data. *Multivariate Behavioral Research*, *53*(4), 453-480. https://doi.org/10.1080/00273171.2018.1454823

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research:Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378. https://doi.org/10.1177/1948550617693063

Fried, E. I., & Cramer, A. O. J. (2017). Moving Forward: Challenges and Directions for Psychopathological Network Theory and Methodology. *Perspectives on Psychological Science*, *12*(6), 999-1020. https://doi.org/10.1177/1745691617705892

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. *Practical assessment, research & evaluation*, *17*(3).

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLOS ONE*, *12*(6), e0174035. https://doi.org/10.1371/journal.pone.0174035

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

Goodman, R., Meltzer, H., & Bailey, V. (1998). The strengths and difficulties questionnaire: A pilot study on the validity of the self-report version. *European Child & Adolescent Psychiatry*, *7*(3), 125-130. https://doi.org/10.1007/s007870050057

Grych, J., Taylor, E., Banyard, V., & Hamby, S. (2020). Applying the dual factor model of mental health to understanding protective factors in adolescence. *American Journal of Orthopsychiatry*, -No Pagination Specified. https://doi.org/10.1037/ort0000449

Hallquist, M. N., Wright, A. G. C., & Molenaar, P. C. M. (2019). Problems with Centrality Measures in Psychopathology Symptom Networks: Why Network Psychometrics Cannot Escape Psychometric Theory. *Multivariate Behavioral Research*, 1-25. https://doi.org/10.1080/00273171.2019.1640103

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

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Humphrey, N. (2013). *Social and emotional learning: A critical appraisal*. SAGE Publications Limited. Hunter, S. C., Houghton, S., & Wood, L. (2015). Positive Mental Well-being in Australian Adolescents: Evaluating the Warwick-Edinburgh Mental Well-being Scale. *The Australian Educational and Developmental Psychologist*, *32*(2), 93-104. https://doi.org/10.1017/edp.2015.12

Johnston, C., & Gowers, S. (2005). Routine Outcome Measurement: A Survey of UK Child and Adolescent Mental Health Services. *Child and Adolescent Mental Health*, *10*(3), 133-139. https://doi.org/doi:10.1111/j.1475-3588.2005.00357.x

Kan, K.-J., van der Maas, H. L. J., & Levine, S. Z. (2019). Extending psychometric network analysis: Empirical evidence against g in favor of mutualism? *Intelligence*, *73*, 52-62. https://doi.org/https://doi.org/10.1016/j.intell.2018.12.004

Kuehner, C. (2017). Why is depression more common among women than among men? *The Lancet Psychiatry*, *4*(2), 146-158. https://doi.org/https://doi.org/10.1016/S2215-0366(16)30263-2

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01398-0

Merikangas, K. R., He, J.-p., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., . . . Swendsen, J. (2010). Lifetime Prevalence of Mental Disorders in U.S. Adolescents: Results from the National

Comorbidity Survey Replication–Adolescent Supplement (NCS-A). *Journal of the American Academy of Child &*

Adolescent Psychiatry, *49*(10), 980-989. https://doi.org/https://doi.org/10.1016/j.jaac.2010.05.017

Moore, S. A., Dowdy, E., Nylund-Gibson, K., & Furlong, M. J. (2019). A latent transition analysis of the longitudinal stability of dual-factor mental health in adolescence. *Journal of School Psychology*, *73*, 56-73. https://doi.org/https://doi.org/10.1016/j.jsp.2019.03.003

Newson, J. J., Hunter, D., & Thiagarajan, T. C. (2020). The Heterogeneity of Mental Health Assessment. *Frontiers in Psychiatry*, *11*(76). https://doi.org/10.3389/fpsyt.2020.00076

Ng Fat, L., Scholes, S., Boniface, S., Mindell, J., & Stewart-Brown, S. (2017). Evaluating and establishing national norms for mental wellbeing using the short Warwick–Edinburgh Mental Well-being Scale (SWEMWBS): findings from the Health Survey for England. *Quality of Life Research*, *26*(5), 1129-1144. https://doi.org/10.1007/s11136-016-1454-8

NHS Digital. (2018). *Mental Health of Children and Young People in England, 2017 Summary of key findings*. https://files.digital.nhs.uk/F6/A5706C/MHCYP%202017%20Summary.pdf

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*(2), 173-182. https://doi.org/10.1038/s41562-018-0506-1

Ortuño-Sierra, J., Chocarro, E., Fonseca-Pedrero, E., Riba, S. S. i., & Muñiz, J. (2015). The assessment of emotional and Behavioural problems: Internal structure of The Strengths and Difficulties Questionnaire. *International Journal of Clinical and Health Psychology*, *15*(3), 265-273. https://doi.org/https://doi.org/10.1016/j.ijchp.2015.05.005

Pan, J., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The Bayesian lasso. *Psychological Methods*, *22*(4), 687-704. https://doi.org/10.1037/met0000112

Patalay, P., & Fitzsimons, E. (2016). Correlates of Mental Illness and Wellbeing in Children: Are They the Same? Results From the UK Millennium Cohort Study. *J Am Acad Child Adolesc Psychiatry*, *55*(9), 771-783. https://doi.org/10.1016/j.jaac.2016.05.019

Perry, D. G., & Pauletti, R. E. (2011). Gender and Adolescent Development. *Journal of Research on Adolescence*, *21*(1), 61-74. https://doi.org/10.1111/j.1532-7795.2010.00715.x

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor Models and Rotations: Exploring the Extent to Which Multidimensional Data Yield Univocal Scale Scores. *Journal of Personality Assessment*, *92*(6), 544-559. https://doi.org/10.1080/00223891.2010.496477

Renshaw, T. L., & Bolognino, S. J. (2017). Psychometrics of the Psychological Wellbeing and Distress Screener: A Brief Measure of Youth's Bidimensional Mental Health. *Assessment for Effective Intervention*, *42*(3), 160-167. https://doi.org/10.1177/1534508416678970

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. *Journal of Personality Assessment*, *98*(3), 223-237. https://doi.org/10.1080/00223891.2015.1089249

Rosseel, Y., Jorgensen, T., Oberski, D., Vanbrabant, L., Savalei, V., Merkle, E., . . . Scharf, F. (2020). *Package 'lavaan'*. https://cran.r-project.org/web/packages/lavaan/lavaan.pdf

Rueger, S. Y., Malecki, C. K., & Demaray, M. K. (2010). Relationship Between Multiple Sources of Perceived Social Support and Psychological and Academic Adjustment in Early Adolescence: Comparisons Across Gender. *Journal of Youth and Adolescence*, *39*(1), 47. https://doi.org/10.1007/s10964-008-9368-6

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J., & Weich, S. (2009). Internal construct validity of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a Rasch analysis using data from the Scottish Health Education Population Survey. *Health and Quality of Life Outcomes*, *7*(1), 15. https://doi.org/10.1186/1477-7525-7-15

Suldo, S., Thalji-Raitano, A., Kiefer, S. M., & Ferron, J. M. (2016). Conceptualizing High School Students' Mental Health Through a Dual-Factor Model. *School Psychology Review*, *45*(4), 434-457. https://doi.org/10.17105/spr45-4.434-457

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., . . . Stewart-Brown, S. (2007). The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health and Quality of Life Outcomes*, *5*(1), 63. https://doi.org/10.1186/1477-7525-5-63

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236-247. https://doi.org/https://doi.org/10.1016/j.ijresmar.2010.02.004

Williams, D. R., & Rast, P. (2019). Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*, *73*(2). https://doi.org/10.1111/bmsp.12173

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1-37. https://doi.org/10.1017/S0140525X20001685

Youth in Mind. (2016). *Scoring the Strengths & Difficulties Questionnaire for age 4-17 or 18+*. Retrieved 07/09/2018 from http://www.sdqinfo.com/py/sdqinfo/b3.py?language=Englishqz(UK)

Table 1. Sample characteristics

Characteristics (%)		Sample 1	Sample 2	National figure 2017/2018 or published norm
Participants for whom demographic data recorded (before exclusions)		95.6	90.4	-
Girls		52.2	52.7	-
Ever eligible for free school meals		34.7	30.6	28.5/28.5
Statement of special needs		10.2	9.6	14.4/14.6
Ethnicity	Asian	9.7	8.2	10.7/11.9
	Black	5.7	5.1	5.6/5.8
	Chinese	.2	.3	.4/.4
	Mixed	3.8	3.4	5.0/5.2
	White	74.1	71.3	75.2/74.2
	Other	1.4	1.5	1.7/1.8
	Unclassified	.8	.6	1.5/1.5
Mean SWEMWBS score (standard deviation)		21.54 (4.49)	21.64 (4.56)	23.57 (3.61)
Emotional symptoms normal %		69.13	69.69	80
Emotional symptoms borderline %		9.59	9.73	10
Emotional symptoms abnormal %		20.31	19.91	10

Note: SWEMWBS = Short Warwick-Edinburgh Mental Well-being Scale.

Table 2.

Descriptive Statistics For items in Each Sample

Item and abbreviated wording	Sample	*M* (SD)	skew
SWEMWBS 1 “feeling optimistic about the future”	S1	3.28 (1.06)	-0.24
	S2	3.29 (1.08)	-0.26
SWEMWBS 2 “ feeling useful”	S1	3.12 (1.03)	-0.13
	S2	3.12 (1.04)	-0.15
SWEMWBS 3 “ feeling relaxed”	S1	3.14 (1.12)	-0.07
	S2	3.14 (1.11)	-0.09
SWEMWBS 4 “ dealing with problems well”	S1	3.23 (1.11)	-0.23
	S2	3.24 (1.11)	-0.25
SWEMWBS 5 “thinking clearly”	S1	3.29 (1.07)	-0.25
	S2	3.34 (1.09)	-0.29
SWEMWBS 6 “feeling close to other people”	S1	3.55 (1.10)	-0.52
	S2	3.57 (1.10)	-0.51
SWEMWBS 7 “able to make up my own mind about things”	S1	3.76 (1.06)	-0.65
	S2	3.77 (1.07)	-0.65
SDQ 13 “often unhappy, down-hearted or tearful”	S1	1.59 (0.71)	0.77
	S2	1.59 (0.70)	0.75
SDQ 3 “headaches, stomach-aches or sickness”	S1	1.78 (0.75)	0.39

	S2	1.77 (0.75)	0.41
	S1	2.03 (0.78)	-0.06
SDQ 8 “worry a lot”	S2	2.04 (0.78)	-0.06
	S1	2.08 (0.76)	-0.14
SDQ 16 “ nervous in new situations/easily lose confidence"	S2	2.08 (0.76)	-0.13
	S1	1.63 (0.72)	0.69
SDQ 24 "I have many fears/easily scared"	S2	1.62 (0.72)	0.70

Note. SWEMWBS = Short Warwick-Edinburgh Mental Well-being Scale; SDQ = Strengths and Difficulties Questionnaire.

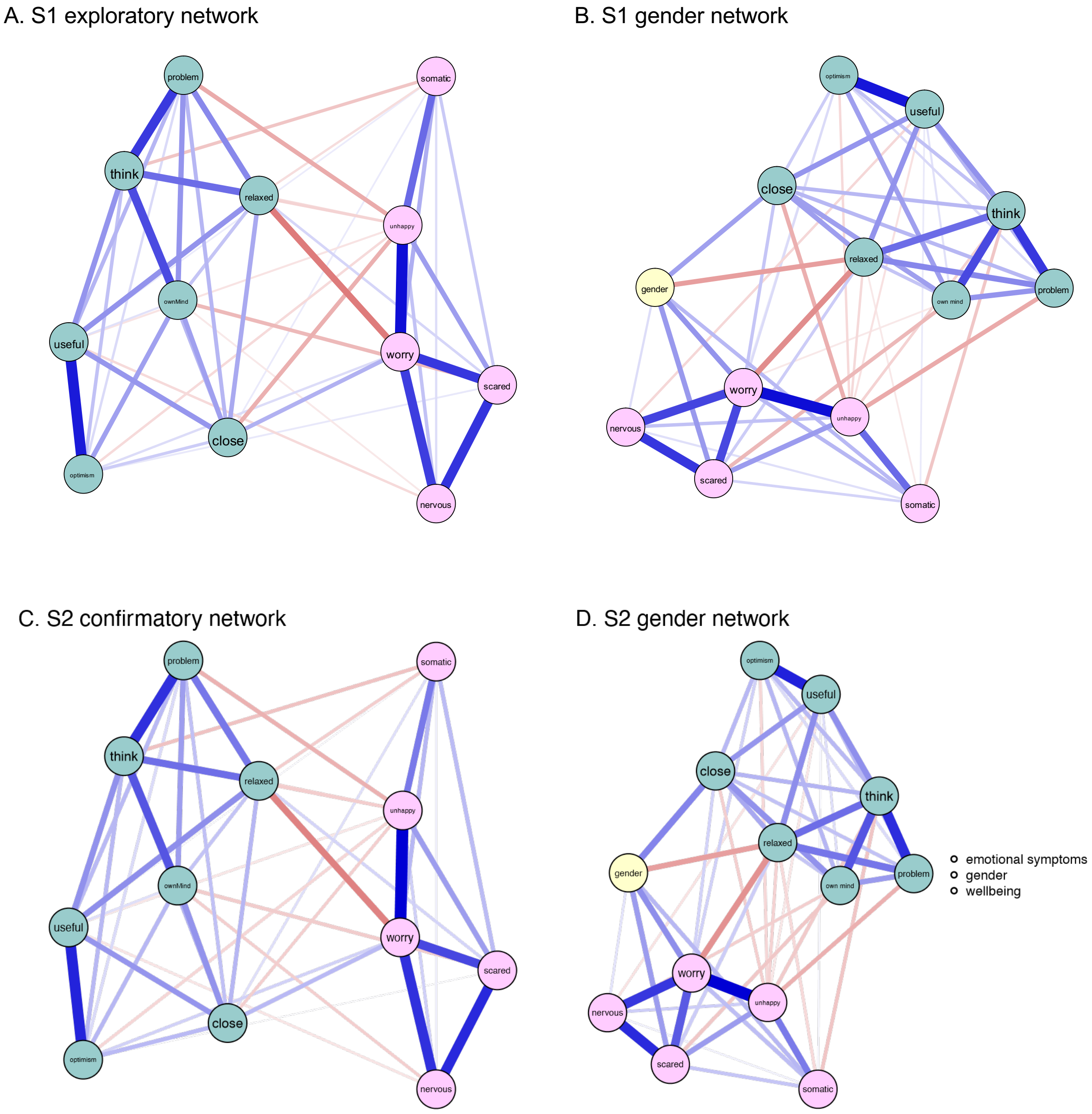
Table 3.

Descriptive statistics of exploratory and confirmatory networks, including by subsection

Network section	Density %	Min	Max	*M*	SD
Whole network	72.72	.03/.005	.33/.35	.12/.12	.08/.08
Within wellbeing	95.24	.04/.04	.31/.30	.14/14	.07/.07
Within emotional symptoms	100	.05/.02	.33/.35	.18/.18	.10/.11
Cross-domain	51.42	.03/.005	.18/.17	.07/.06	.04/.04

Note. Density represents the percentage of edges retained out of all possible relationships, given the number of variables under consideration. Except for density, where results are the same for both models, exploratory and confirmatory results are shown in the following format within each cell exploratory/confirmatory. As can be seen from Figure 1, both positive and negative edges were returned, but absolute values are shown in this Table to give better insight into ranges.

Figure 1. Exploratory, confirmatory and known-groups validity networks



A. S1 exploratory network

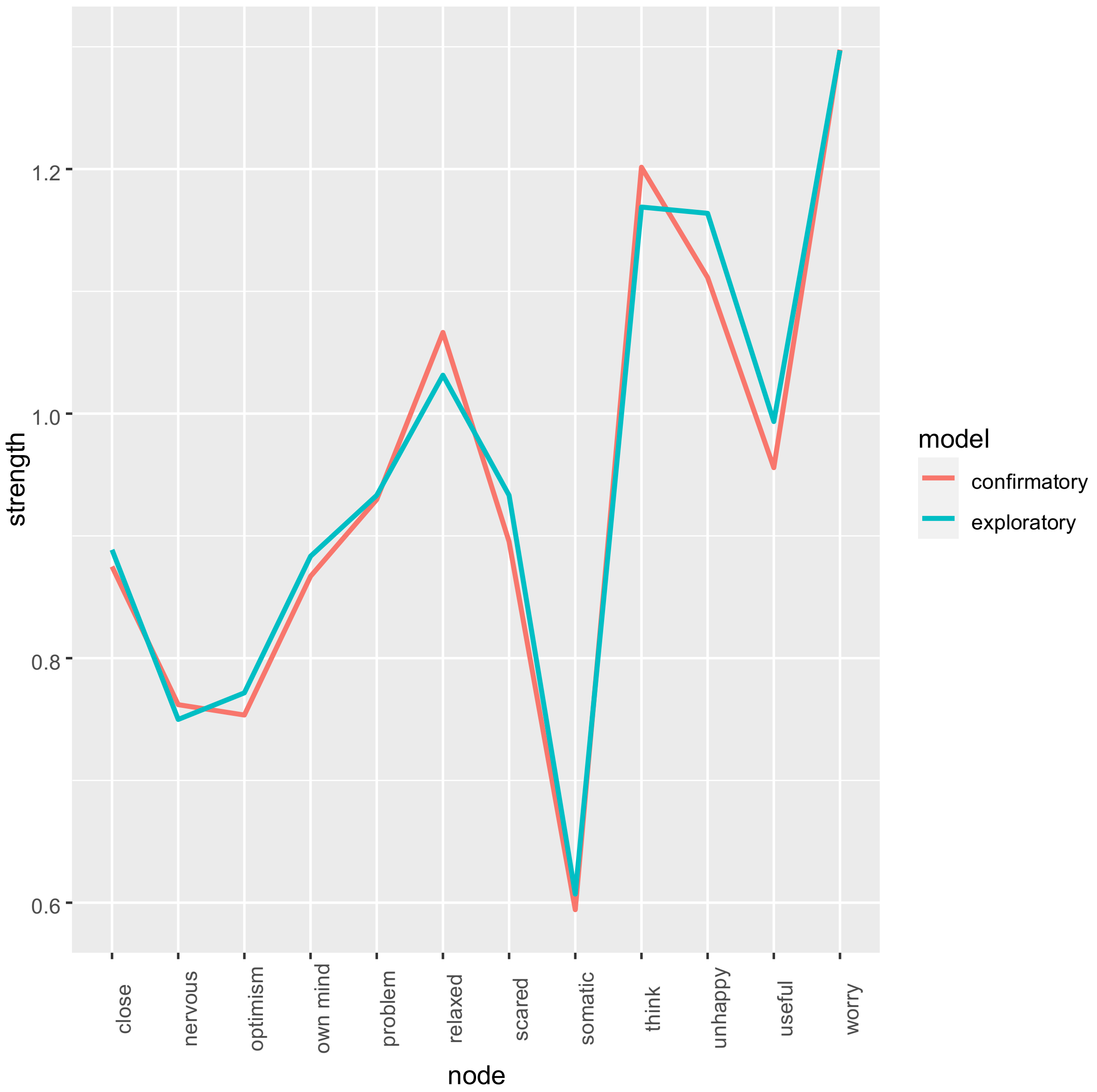
B. S1 gender network

C. S2 confirmatory network

D. S2 gender network

Figure 2 Node strength for exploratory and confirmatory networks

Node strength


strength
1.2
1.0
0.8
0.6
close
nervous
optimism
own mind
problem
relaxed
scared
somatic
think
unhappy
useful
worry
node
model
confirmatory
exploratory

Appendix 2: Supplementary Materials for Paper 1

S1 Appendix: Items of Me and My School and Child Outcome Rating Scale Questionnaires

Each instrument is reproduced here with the items listed in the order presented in the current study.

Me and My School

I feel lonely

I am unhappy

Nobody likes me

I cry a lot

I worry when I am at school

I worry a lot

I have problems sleeping

I wake up in the night

I am shy

I feel scared

I get very angry

I lose my temper

I do things to hurt people

I am calm

I hit out when I am
angry I break things on
purpose

Child Outcome Rating Scale

Me

(How am I doing?)

Family

(How are things in my family?)

School

(How am I doing at school?)

Everything

(How is everything going?)

Mplus Code Examples Paper 1

CORS CFA

```
Title: Cors Cfa step 1
Data: File= 'recoded trimmed dataset.dat';
    Format= 2f12.0, f1.0, 18f12.0, 4f8.2;
```

```
Variable: Names= pupilid, schoolid, gender, mams1-mams16, ever6, everall,
cors1-cors4;
    Usevariables= cors1, cors2, cors3, cors4;
    Missing= All(-99);
Cluster= schoolid;


Analysis: Type= complex;
    Estimator= MLR;


Model: f1 by cors1-cors4; cors1
with cors3@0;




Output: sampstat stand mod;
```

Correlated factors model

```
Title: complex cfa (correlated factors) HS2 with cors and mams, minus mams9
with cors1/3 error correlation


Data: File= 'recoded trimmed dataset.dat';
    Format= 2f12.0, f1.0, 18f12.0, 4f8.2;


Variable: Names= pupilid, schoolid, gender, mams1-mams16, ever6, everall,
cors1-cors4;
    Usevariables= mams1-mams8, mams10-mams16 cors1, cors2, cors3, cors4;
    Categorical= mams1-mams8, mams10-mams16;
    Missing= All(-99);

     cluster=

schoolid;



Analysis: Type= complex;

Estimator= WLSMV;



Model: F1 by mams1-mams8 mams10;

F2 by mams11-mams16; F3

by cors1-cors4; cors1

with cors3; mams3 with

mams1; mams5 with

mams6; mams8 with

mams7;



Output: sampstat stand mod;
```

Bifactor

```
Title: bifactor complete mental health CFA without mams9, cors1/3 error
correlated


Data: File= 'recoded trimmed dataset.dat';
    Format= 2f12.0, f1.0, 18f12.0, 4f8.2;


Variable: Names= pupilid, schoolid, gender, mams1-mams16, ever6, everall,
cors1-cors4;
    Usevariables= mams1-mams8, mams10-mams16, cors1, cors2, cors3, cors4;
    Categorical=  mams1-mams8, mams10-mams16;
    Missing= All(-99);


    cluster= schoolid;


Analysis: Type= complex;
Estimator= WLSMV;


Model: F1 by mams1-mams8 mams10;
```

```
F2 by mams11-mams16;
F3 by cors1-cors4;
Fg by mams1-mams8 mams10-mams16 cors1-cors4;
F1 with F2@0; F1 with F3@0; F1 with Fg@0; F2 with F3@0; F2 with Fg@0; F3
with Fg@0; cors1 with cors3; mams3 with mams1; mams5 with mams6; mams8
with mams7;


Output: sampstat stand mod;
```

ECV

```
Title: ECV bifactor complete mental health CFA without mams9, cors1/3 error
correlated


Data: File= 'recoded trimmed dataset.dat';
    Format= 2f12.0, f1.0, 18f12.0, 4f8.2;


Variable: Names= pupilid, schoolid, gender, mams1-mams16, ever6, everall,
cors1-cors4;
    Usevariables= mams1-mams8, mams10-mams16, cors1, cors2, cors3, cors4;
    Categorical=  mams1-mams8, mams10-mams16;
    Missing= All(-99);


    cluster= schoolid;


Analysis: Type= complex;
Estimator= WLSMV;
Model:
  !FACTOR 1   F1 by
mams1* (lm1)   mams2-
mams8 (lm2-lm8)
mams10 (lm10);


  !FACTOR 2
  F2 by mams11* (lm11)   mams12-
mams16 (lm12-lm16);


  !FACTOR 3   F3 by
cors1* (lc1)   cors2-
cors4 (lc2-lc4);




  !BIFACTOR GENERAL FACTOR
Fg by mams1* (lgm1)   mams2-
mams8 (lgm2-lgm8)   mams10-
mams16 (lgm10-lgm16)   cors1-
cors4 (lgc1-lgc4);


  F1-Fg@1;
  F1 with F2@0; F1 with F3@0; F1 with Fg@0; F2 with F3@0; F2 with Fg@0; F3
with Fg@0;
  cors1 with cors3;
mams3 with mams1;   mams5
with mams6;   mams8 with
mams7;


  MODEL CONSTRAINT:

  NEW (RLM1 RLM2 RLM3 RLM4 RLM5 RLM6 RLM7 RLM8 RLM10
       RLM11 RLM12 RLM13 RLM14 RLM15 RLM16
       RLC1 RLC2 RLC3 RLC4
       RLGM1 RLGM2 RLGM3 RLGM4 RLGM5 RLGM6 RLGM7 RLGM8 RLGM10
       RLGM11 RLGM12 RLGM13 RLGM14 RLGM15 RLGM16
       RLGC1 RLGC2 RLGC3 RLGC4
       VGF VF1 VF2 VF3 ECV);

  !INDIVIDUAL FACTORS
  RLM1 = LM1**2;
  RLM2 = LM2**2;
  RLM3 = LM3**2;
  RLM4 = LM4**2;
  RLM5 = LM5**2;
  RLM6 = LM6**2;
  RLM7 = LM7**2;
  RLM8 = LM8**2;
  RLM10 = LM10**2;


  RLM11 = LM11**2;
  RLM12 = LM12**2;
  RLM13 = LM13**2;
  RLM14 = LM14**2;
  RLM15 = LM15**2;
  RLM16 = LM16**2;


  RLC1 = LC1**2;
  RLC2 = LC2**2;
  RLC3 = LC3**2;
  RLC4 = LC4**2;

  !GENERAL FACTOR


  RLGM1 = LGM1**2;
```

```
RLGM2 = LGM2**2;
RLGM3 = LGM3**2;
RLGM4 = LGM4**2;
RLGM5 = LGM5**2;
RLGM6 = LGM6**2;
RLGM7 = LGM7**2;
RLGM8 = LGM8**2;
RLGM10 = LGM10**2;



RLGM11 = LGM11**2;
RLGM12 = LGM12**2;
RLGM13 = LGM13**2;
RLGM14 = LGM14**2;
RLGM15 = LGM15**2;
RLGM16 = LGM16**2;



RLGC1 = LGC1**2;
RLGC2 = LGC2**2;
RLGC3 = LGC3**2;
RLGC4 = LGC4**2;



VGF = RLGM1+RLGM2+RLGM3+RLGM4+RLGM5+RLGM6+RLGM7+RLGM8+RLGM10+
RLGM11+RLGM12+RLGM13+RLGM14+RLGM15+RLGM16+
RLGC1+RLGC2+RLGC3+RLGC4;


VF1 = RLM1+RLM2+RLM3+RLM4+RLM5+RLM6+RLM7+RLM8+RLM10;


VF2 = RLM11+RLM12+RLM13+RLM14+RLM15+RLM16;


VF3 = RLGC1+RLGC2+RLGC3+RLGC4;



ECV = VGF/(VGF+VF1+VF2+VF3);




Output: sampstat stand mod;
```

S-1 internalizing

```
Title: S-1 internalizing complete mental health CFA without mams9, cors1/3
error correlated


Data: File= 'recoded trimmed dataset.dat';
    Format= 2f12.0, f1.0, 18f12.0, 4f8.2;


Variable: Names= pupilid, schoolid, gender, mams1-mams16, ever6, everall,
cors1-cors4;
    Usevariables= mams1-mams8, mams10-mams16, cors1, cors2, cors3, cors4;
    Categorical=  mams1-mams8, mams10-mams16;
    Missing= All(-99);


    cluster= schoolid;     !
ever6 (0=never, 1= ever);


Analysis: Type= complex;
Estimator= WLSMV;
Model:
!F1 by mams1-mams8 mams10;
F2 by mams11-mams16;
F3 by cors1-cors4;
Fg by mams1-mams8 mams10-mams16 cors1-cors4;
 F2 with F3@0; F2 with Fg@0; F3 with Fg@0;
cors1 with cors3; mams3 with mams1; mams5
with mams6; mams8 with mams7;
Output: sampstat stand mod;
```

Measurement invariance

Boys baseline configural

```
Title: complex cfa (correlated factors) HS2 with cors and mams, minus mams9
with cors1/3 error correlation


Data: File= 'recoded trimmed dataset.dat';
    Format= 2f12.0, f1.0, 18f12.0, 4f8.2;


Variable: Names= pupilid, schoolid, gender, mams1-mams16, ever6, everall,
cors1-cors4;
    Usevariables= mams1-mams8, mams10-mams16 cors1, cors2, cors3, cors4;
    Categorical= mams1-mams8, mams10-mams16;
Missing= All(-99);      subpopulation=
gender eq 1;


cluster= schoolid;


Analysis: Type= complex;
Estimator= WLSMV;
```

```
Model: F1 by mams1-mams8 mams10;
F2 by mams11-mams16; F3
by cors1-cors4; cors1
with cors3; mams3 with
mams1; mams5 with
mams6; mams8 with
mams7;


Output: sampstat stand mod;
```

Correlated factors configural

```
Title: bifactor complete mental health CFA without mams9, cors1/3 error
correlated-
  configural invariance following Uni Kentucky example but reference
indicator method


  Data: File= 'recoded trimmed dataset.dat';
      Format= 2f12.0, f1.0, 18f12.0, 4f8.2;


  Variable: Names= pupilid, schoolid, gender, mams1-mams16, ever6, everall,
cors1-cors4;
      Usevariables= mams1-mams8, mams10-mams16, cors1, cors2, cors3, cors4;
      Categorical=  mams1-mams8, mams10-mams16;
      Missing= All(-99);


      cluster= schoolid;
        grouping= gender (1= boys 2= girls);
  Analysis: Type= complex;
Estimator= wlsmv;   parameterization=
theta; Model:
  !baseline
  !factor loadings ALL FREELY estimated
F1 by    mams7*(lm7)            mams2*
(lm2)               mams3* (lm3)
  mams4* (lm4)            mams5* (lm5)
        mams6* (lm6)
mams1* (lm1)              mams8* (lm8)
        mams10* (lm10);


      F2 by    mams11* (lm11)
  mams12* (lm12)                mams13*
(lm13)              mams14* (lm14)
        mams15* (lm15)
mams16* (lm16);


      F3 by    cors1* (lc1)
         cors2* (lc2)
cors3* (lc3)
cors4* (lc4);     mams3 with
```

```
mams1; mams5 with mams6; mams8
with mams7; cors1 with cors3;



  !Free factor variances
F1-F3@1;




  !Item thresholds/intercepts all free


  [mams1$1*];
  [mams1$2*];
  [mams2$1*];
  [mams2$2*];
  [mams3$1*];
  [mams3$2*];
  [mams4$1*];
  [mams4$2*];
  [mams5$1*];
  [mams5$2*];
  [mams6$1*];
  [mams6$2*];
  [mams7$1*];
  [mams7$2*];
  [mams8$1*];
  [mams8$2*];
  [mams10$1*];
  [mams10$2*];
  [mams11$1*];
  [mams11$2*];
  [mams12$1*];
  [mams12$2*];
  [mams13$1*];
  [mams13$2*];
  [mams14$1*];
  [mams14$2*];
  [mams15$1*];
  [mams15$2*];
  [mams16$1*];
  [mams16$2*];
  [cors1*]
  [cors2*]
  [cors3*]
  [cors4*]


  !Factor means fixed@0 for identification
       [F1@0 F2@0 F3@0];
  !Item residual variances all fixed@1
  mams1@1 mams2@1 mams3@1 mams4@1 mams5@1 mams6@1 mams7@1 mams8@1 mams10@1
mams11@1 mams12@1 mams13@1 mams14@1 mams15@1 mams16@1   cors1@1 cors2@1
cors3@1 cors4@1;


    model
girls:


  !factor loadings ALL FREELY estimated
        F1 by    mams7*
             mams2*
  mams3*
mams4*              mams5*
        mams6*
mams1*              mams8*
        mams10*;

F2 by    mams11*
             mams12*
  mams13*                mams14*
        mams15*
mams16* ;


      F3 by    cors1*
cors2*               cors3*
         cors4*;


      mams3 with mams1;
mams5 with mams6; mams8
with mams7; cors1 with
cors3;


!Free factor variances
F1-F3@1;








  !Item thresholds/intercepts all free


  [mams1$1*];
  [mams1$2*];
  [mams2$1*];
  [mams2$2*];
  [mams3$1*];
  [mams3$2*];
  [mams4$1*];
  [mams4$2*];
  [mams5$1*];
  [mams5$2*];
  [mams6$1*];
  [mams6$2*];
  [mams7$1*];
  [mams7$2*];
  [mams8$1*];
  [mams8$2*];
  [mams10$1*];
  [mams10$2*];
  [mams11$1*];
  [mams11$2*];
  [mams12$1*];
  [mams12$2*];
  [mams13$1*];
  [mams13$2*];
  [mams14$1*];
  [mams14$2*];
  [mams15$1*];
  [mams15$2*];
  [mams16$1*];
  [mams16$2*];
  [cors1*];
  [cors2*];
  [cors3*];
  [cors4*];


  !Factor means fixed@0 for identification
        [F1@0 F2@0 F3@0];




  !Item residual variances all fixed@1
  mams1@1 mams2@1 mams3@1 mams4@1 mams5@1 mams6@1 mams7@1 mams8@1 mams10@1
mams11@1 mams12@1 mams13@1 mams14@1 mams15@1 mams16@1   cors1@1 cors2@1
cors3@1;

  savedata: difftest is gender_configural.dat;


  Output: sampstat stand mod;
```

Correlated factors scalar

```
Title: bifactor complete mental health CFA without mams9, cors1/3 error
correlated-
  configural invariance following Uni Kentucky example but reference
indicator method


  Data: File= 'recoded trimmed dataset.dat';
      Format= 2f12.0, f1.0, 18f12.0, 4f8.2;


  Variable: Names= pupilid, schoolid, gender, mams1-mams16, ever6, everall,
cors1-cors4;
```

```
      Usevariables= mams1-mams8, mams10-mams16, cors1, cors2, cors3, cors4;
      Categorical=  mams1-mams8, mams10-mams16;
      Missing= All(-99);


      cluster= schoolid;
        grouping= gender (1= boys 2= girls);
  Analysis: Type= complex;
Estimator= wlsmv;
parameterization= theta; difftest
is gender_configural.dat;
  Model:
  !baseline
  !factor loadings ALL FREELY estimated
F1 by    mams7(lm7)
              mams2* (lm2)
  mams3* (lm3)
mams4* (lm4)              mams5*
(lm5)               mams6* (lm6)
        mams1* (lm1)
mams8* (lm8)              mams10*
(lm10);


        F2 by    mams11 (lm11)
              mams12* (lm12)
  mams13* (lm13)                mams14*
(lm14)              mams15* (lm15)
                 mams16* (lm16);


      F3 by    cors1 (lc1)
         cors2* (lc2)
cors3* (lc3)               cors4*
(lc4);
      mams3 with mams1;
mams5 with mams6; mams8
with mams7; cors1 with
cors3;




  !Free factor variances
F1-F3*;




  !Item thresholds/intercepts all free


  [mams1$1*];
  [mams1$2*];
  [mams2$1*];
  [mams2$2*];
  [mams3$1*];
  [mams3$2*];
  [mams4$1*];
  [mams4$2*];
  [mams5$1*];
  [mams5$2*];
  [mams6$1*];
  [mams6$2*];
  [mams7$1*];
  [mams7$2*];
  [mams8$1*];
  [mams8$2*];
  [mams10$1*];
  [mams10$2*];
  [mams11$1*];
  [mams11$2*];
  [mams12$1*];
  [mams12$2*];
  [mams13$1*];
  [mams13$2*];
  [mams14$1*];
  [mams14$2*];
  [mams15$1*];
  [mams15$2*];
  [mams16$1*];
  [mams16$2*];
  [cors1*];
  [cors2*];
  [cors3*];
  [cors4*];


  !Factor means fixed@0 for identification
        [F1@0 F2@0 F3@0];






  !Item residual variances all fixed@1
  mams1@1 mams2@1 mams3@1 mams4@1 mams5@1 mams6@1 mams7@1 mams8@1 mams10@1
mams11@1 mams12@1 mams13@1 mams14@1 mams15@1 mams16@1   cors1@1 cors2@1
cors3@1 cors4@1;


    model
girls:


  !factor loadings ALL FREELY estimated
F1 by    mams7@1 (lm7)
mams2* (lm2)             mams3* (lm3)
        mams4* (lm4)
mams5* (lm5)             mams6* (lm6)
        mams1* (lm1)
mams8* (lm8)             mams10*
(lm10);


      F2 by    mams11@1 (lm11)
  mams12* (lm12)                mams13*
```

```
(lm13)              mams14* (lm14)
        mams15* (lm15)
                 mams16* (lm16);


      F3 by    cors1@1 (lc1)
cors2* (lc2)              cors3*
(lc3)
                 cors4* (lc4);


      mams3 with mams1;
mams5 with mams6; mams8
with mams7; cors1 with
cors3;


!Free factor variances
F1-F3*;


  !Factor means fixed@0 for identification
        [F1* F2* F3*];


  !Item residual variances all fixed@1
  mams1@1 mams2@1 mams3@1 mams4@1 mams5@1 mams6@1 mams7@1 mams8@1 mams10@1
mams11@1 mams12@1 mams13@1 mams14@1 mams15@1 mams16@1   cors1@1 cors2@1
cors3@1;


  Output: sampstat stand mod;
```

Appendix 3: Supplementary Materials for Paper 3

Supplemental Material for: Age appropriateness of the self-report Strengths and Difficulties Questionnaire

Supplemental Table S1.

Exploratory Structural Equation Model Parameters

Theoretical dimension	Item/factor	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
emotional problems	3	*0.104*	***0.363***	*0.057*	*0.16*	*0.029*
	8	*-0.029*	***0.725***	*0.052*	*-0.028*	*0.034*
	13	*-0.032*	***0.523***	*-0.045*	*0.193*	*0.224*
	16	*0.077*	***0.579***	*-0.037*	*-0.109*	0.025
	24	-0.001	***0.536***	*0.038*	-0.009	*0.123*
conduct problems	5	*0.2*	*0.18*	-0.024	***0.431***	-0.018
	7	*0.112*	*-0.024*	***-0.339***	***0.314***	*-0.228*
	12	-0.004	-0.007	-0.012	***0.613***	0.015
	18	*0.149*	*0.03*	*-0.03*	***0.405***	*0.197*
	22	-0.009	*-0.037*	*-0.022*	***0.401***	*0.142*
hyperactivity	2	***0.724***	*-0.081*	0.009	*-0.036*	*0.065*
	10	***0.768***	*-0.036*	*0.04*	-0.011	*0.103*
	15	***0.553***	*0.149*	*-0.067*	*0.11*	-0.075
	21	*0.193*	*0.082*	***-0.363***	*0.14*	*-0.183*
	25	*0.294*	*0.108*	***-0.405***	0.058	*-0.19*
peer problems	6	0.031	*0.103*	*-0.1*	*0.039*	***0.493***
	11	*-0.095*	0.005	***-0.342***	0.003	***0.316***

	14	*0.019*	*0.086*	***-0.419***	*-0.181*	***0.359***
	19	*0.041*	*0.131*	0.027	*0.193*	***0.477***
	23	*0.06*	*0.06*	*0.063*	*0.143*	***0.377***
prosocial	1	0.008	*0.151*	***0.527***	*-0.218*	-0.03
	4	0.005	*0.124*	***0.454***	*0.044*	*-0.103*
	9	0.017	*0.157*	***0.623***	-0.005	-0.036
	17	*0.058*	*0.107*	***0.476***	*-0.109*	*-0.031*
	20	-0.009	-0.028	***0.569***	*0.088*	*0.204*
	factor 1	-	*0.351*	*-0.226*	*0.485*	-0.007
	factor 2	-	-	-0.028	*0.169*	*0.382*
	factor 3	-	-	-	*-0.366*	0.011
	factor 4	-	-	-	-	0.104

Note. Robust maximum likelihood was used. Bolded coefficients are higher than .30. Italicized coefficients are significant at p <. 01. Underlined coefficients are secondary loadings with a discrepancy <.30 compared to the highest loading.

Mplus Code examples from Paper 3

5-factor, MLR

```
  Title: sdq 5 factor HS T1;
   data:    file=
'june.sdq.dat';
format= f5.0, 26f4.0;

variable:
   names= school, yearg, sdq1, sdq2, sdq3, sdq4, sdq5, sdq6, sdq7,
sdq8, sdq9,   sdq10, sdq11, sdq12, sdq13, sdq14, sdq15, sdq16, sdq17,
sdq18,   sdq19, sdq20, sdq21, sdq22, sdq23, sdq24, sdq25;
   usevariables=  sdq1-
sdq25;


  Missing= All(-999);



  Analysis:
  Estimator= MLR;


!type = complex;
  Model:
  ES by sdq3 sdq8 sdq13  sdq16  sdq24;
  CP by sdq5 sdq7 sdq12 sdq18 sdq22;
  HI by sdq2 sdq10 sdq15 sdq21 sdq25;
  PP by sdq6 sdq11 sdq14 sdq19 sdq23;
  PS by sdq1 sdq4 sdq9 sdq17 sdq20;
  Output: sampstat stand mod residual;
```

5-factor WLSMV sensitivity

```
Title: sdq 5 factor wlsmv HS T1;
   data:    file=
'june.sdq.dat';
format= f5.0, 26f4.0;

variable:
   names= school, yearg, sdq1, sdq2, sdq3, sdq4, sdq5, sdq6, sdq7,
sdq8, sdq9,   sdq10, sdq11, sdq12, sdq13, sdq14, sdq15, sdq16, sdq17,
sdq18,   sdq19, sdq20, sdq21, sdq22, sdq23, sdq24, sdq25;
   usevariables=  sdq1-
sdq25;
   categorical =sdq1, sdq2, sdq3, sdq4, sdq5, sdq6, sdq7, sdq8,
sdq9,   sdq10, sdq11, sdq12, sdq13, sdq14, sdq15, sdq16, sdq17,
sdq18,   sdq19, sdq20, sdq21, sdq22, sdq23, sdq24, sdq25;




  Missing= All(-999);
!cluster = school;
```

```
  Analysis:
  Estimator= wlsmv;
  !type = complex;


  Model:
  ES by sdq3 sdq8 sdq13  sdq16  sdq24;
  CP by sdq5 sdq7 sdq12 sdq18 sdq22;
  HI by sdq2 sdq10 sdq15 sdq21 sdq25;
  PP by sdq6 sdq11 sdq14 sdq19 sdq23;
  PS by sdq1 sdq4 sdq9 sdq17 sdq20;




  Output: sampstat stand mod;
```

Invariance models examples

Year 7 baseline

```
Title: sdq y7bl esem;
   data:    file=
'june.sdq.dat';
format= f5.0, 26f4.0;

variable:
   names= school, yearg, sdq1, sdq2, sdq3, sdq4, sdq5, sdq6, sdq7,
sdq8, sdq9,   sdq10, sdq11, sdq12, sdq13, sdq14, sdq15, sdq16, sdq17,
sdq18,   sdq19, sdq20, sdq21, sdq22, sdq23, sdq24, sdq25;
   usevariables=  sdq1-
sdq25;
```

```
  Missing= All(-999);
   useobservations= yearg
==7;
  Analysis:


  Estimator= MLR;
rotation= geomin;


  Model:
 F1-F5 by sdq1-sdq25 (*1);




  Output: sampstat stand mod tech1 residual;
```

Invariance models

```
Title: sdq esem auto invariance HS T1;
     data:      file=
'june.sdq.dat';
format= f5.0, 26f4.0;

variable:
     names= school, yearg, sdq1, sdq2, sdq3, sdq4, sdq5, sdq6, sdq7, sdq8,
sdq9,     sdq10, sdq11, sdq12, sdq13, sdq14, sdq15, sdq16, sdq17, sdq18,
sdq19, sdq20, sdq21, sdq22, sdq23, sdq24, sdq25;
     usevariables=  sdq1-
sdq25;
```

```
    Missing= All(-999);
     grouping= yearg (7= y7 9=
y9);


    Analysis:


    Estimator= MLR;     rotation=
geomin;     model = configural metric
scalar;


    Model:
   F1-F5 by sdq1-sdq25 (*1);


   Output: sampstat stand mod tech1 residual;
```