Canadian Psychiatric Association
Association des psychiatres du Canada

Regular Article

The Canadian Journal of Psychiatry / La Revue Canadienne de Psychiatrie
2022, Vol. 67(4) 306-315

DOI: 10.1177/07067437211037125
TheCJP.ca | LaRCP.ca
SAGE

The 25-Item Ontario Child Health Study Emotional Behavioural Scales-Brief Version (OCHS-EBS-B): Test-Retest Reliability and Construct Validity When Used as Categorical Measures

Échelles comportementales émotionnelles en 25 items de l'Étude sur la santé des enfants de l'Ontario, version abrégée (OCHS-EBS-B) : fiabilité test-retest et validité du construit lorsqu'elles servent de mesures catégoriques.

Michael H. Boyle, PhD[1], Laura Duncan, PhD[1], Li Wang, PhD[1], and Katholiki Georgiades, PhD[1]

Abstract

Objective: Child and youth mental health problems are often assessed by parent self-completed checklists that produce dimensional scale scores. When converted to binary ratings of disorder, little is known about their psychometric properties in relation to classifications based on lay-administered structured diagnostic interviews. In addition to estimating agreement, our objective is to test for statistical equivalence in the test-retest reliability and construct validity of two instruments used to classify child emotional, behavioural, and attentional disorders: the 25-item, parent completed Ontario Child Health Study Emotional Behavioural Scales-Brief Version (OCHS-EBS-B) and the Mini International Neuropsychiatric Interview for Children and Adolescents-parent version (MINI-KID-P).

Methods: This study draws on independent samples ($n = 452$) and uses the confidence interval approach to test for statistical equivalence. Reliability is based on kappa (κ). Construct validity is based on standardized beta coefficients (β) estimated in structural equation models.

Results: The average differences between the MINI-KID-P and OCHS-EBS-B in κ and β were -0.022 and -0.020, respectively. However, in both instances, criteria for statistical equivalence were met in only 5 of 12 comparisons. Based on κ, between-instrument agreement on the classifications of disorder went from 0.481 (attentional disorder) to 0.721 (emotional disorder) but were substantially higher (0.731 to 0.895, respectively) when corrected for attenuation due to measurement error.

Conclusions: Although falling short of equivalence, the results suggest on balance that the reliability and validity of the two instruments for classifying child psychiatric disorder assessed by parents are highly comparable. This conclusion is supported by the high levels of agreement between the instruments after correcting for attenuation due to measurement error.

Résumé

Objectif : Les problèmes de santé mentale des enfants e des jeunes sont souvent évalués par des listes de vérification remplies par les parents qui produisent des scores dimensionnels aux échelles. Quand elles sont converties à des cotes binaires du

[1] Offord Centre for Child Studies and Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario

Corresponding Author:
Laura Duncan, PhD, Assistant Professor, McMaster University, 1280 Main St W, MIP 201A, Hamilton, Ontario, Canada L8S 4K1.
Email: duncanlj@mcmaster.ca

trouble, elles révèlent très peu de leurs propriétés psychométriques en relation aux classifications basées sur des entrevues diagnostiques structurées (EDS) administrées par un profane. Outre un accord d'évaluation, notre objectif est de vérifier l'équivalence statistique de la fiabilité test-retest et de la validité du construit des deux instruments servant à classifier les troubles émotionnels, comportementaux et de l'attention chez l'enfant : les échelles comportementales émotionnelles en 25 items de l'Étude sur la santé des enfants de l'Ontario remplies par les parents, version abrégée (OCHS-EBS-B) et le Mini-entretien neuropsychiatrique international pour enfants et adolescents (MINI Kid) – version des parents (MINI-KID-P).

Méthodes : La présente étude utilise des échantillons indépendants (n = 452) et l'approche de l'intervalle de confiance pour vérifier l'équivalence statistique. La fiabilité est basée sur kappa (κ). La validité du construit repose sur les coefficients beta normalisés (β) estimés dans des modèles d'équation structurelle.

Résultats : Les différences moyennes entre le MINI-KID-P et l'OCHS-EBS-B dans κ et β étaient de -0,022 et -0,020, respectivement. Cependant, dans les deux cas, les critères de l'équivalence statistique n'étaient satisfaits que dans 5 comparaisons sur 12. Selon κ, l'accord entre instruments sur les classifications du trouble passait de 0,481 (trouble d'attention) à 0,721 (trouble émotionnel) mais était substantiellement plus élevé (0,731 à 0,895, respectivement) avec des corrections de réduction en raison d'erreurs de mesure.

Conclusions : Malgré un déficit d'équivalence, les résultats suggèrent dans l'ensemble que la fiabilité et la validité des deux instruments qui classifient le trouble psychiatrique pédiatrique évalué par les parents sont très comparables. Cette conclusion est soutenue par les niveaux élevés d'accord entre les instruments après correction de réduction en raison d'une erreur de mesure.

Keywords
problem checklist, structured diagnostic interview, measurement, structural equation modelling, validity, reliability, child psychiatric disorder

Introduction

In the 1980s, concern about the reliability of psychiatric diagnoses based on clinical judgement led to the development of structured diagnostic interviews (SDIs) to classify child and youth (herein child/ren) psychiatric disorders as present or absent. Accompanying the development of SDIs was a debate on the nature of child psychiatric disorder: was it a true class in the taxonomic sense or an underlying continuum?[1,2] No evidence has emerged to indicate that common disorders such as generalized anxiety disorder (GAD) or major depressive disorder (MDD), conduct disorder (CD) or oppositional defiant disorder (ODD), and attention-deficit hyperactivity disorder (ADHD) represent true classes. Angold et al.[3] reported large differences in prevalence for the same individuals across SDIs using identical diagnostic criteria, indicating that classifications of child disorder depend on the way instrument developers convert symptoms into classifications. Despite the challenges of classifying disorder, they are needed by clinicians for setting treatment priorities, administrators for making resource allocation decisions, and researchers for conducting epidemiological studies.[4]

Pickles and Angold[1] have called for instruments that can measure child psychiatric disorder dimensionally and classify these disorders as present or absent. Self-completed problem checklists have provided dimensional measures of child psychiatric disorder for decades and exhibit many practical advantages; they are brief, simple, and inexpensive to implement; pose little burden to respondents; and collect information on all symptoms. Although dimensional measures can be converted into classifications of disorder by choosing cut points or thresholds, they are used rarely in this way because SDIs are considered the "gold standard" for doing so.[5,6] In contrast to checklists, SDIs are expensive to implement and burdensome to respondents, making them difficult to use in either community child mental health agencies faced with limited budgets[7] or epidemiological studies faced with increasing nonresponse.[8]

There are theoretical reasons[4] and empirical evidence for expecting classifications of disorders derived from problem checklists to be as reliable and valid as classifications obtained using SDIs. The overall test-retest reliability of SDIs is modest ($\kappa = 0.58$), highly variable across studies,[9] and within reach of scale scores converted to classifications.[10] Furthermore, past empirical studies have been unable to identify validity differences between the two approaches.[11-14]

The objective of this paper is to determine if scale scores from the 25-item, parent completed Ontario Child Health Study Emotional Behavioural Scales-Brief Version (OCHS-EBS-B), converted to binary classifications of these problems, are as reliable and valid as corresponding disorders classified by the Mini International Neuropsychiatric Interview for Children and Adolescents-parent version (MINI-KID-P). In this study, we focus on lay-administered SDIs used in epidemiological studies and clinical research to classify psychiatric disorder as present or absent. This focus excludes the important role that SDIs can play when used by clinicians as aids to the diagnostic process. We hypothesize

that (1) between-instrument differences in test-retest reliability and construct validity will be small and nonsignificant (statistically equivalent); or, and (2) agreement between instruments in the classifications of disorder will be substantial after correcting for attenuation due to measurement error.

Table 1. Study and Sample Characteristics.

Study	A	B	Total	
Study period	2011-2013	2014-2015		
Sample size	284	180	464	χ^2
(% response)	(~17%)	(~50%)		
Sample source				
(1) General population (%)	8 Schools (65%)	CCTB		
(2) Outpatient clinic (%)	2 Clinics (35%)			
Sample characteristics				
n for analysis	277	175	452	
% Male	43.0	51.4	46.2	3.1
% 12–18	85.9	50.9	72.3	65.9‡
% Single parent	29.2	17.1	24.6	8.5†
% Receiving social assistance	13.5	5.7	10.5	6.8†
Disorders				
OCHS-EBS-B				
% Emotional	32.1	4.6	21.5	48.3‡
% Behavioural	26.0	4.0	17.5	36.0‡
% Attentional	9.0	0.6	5.8	14.1‡
MINI-KID-P				
% Emotional	26.7	4.0	17.9	37.6†
% Behavioural	25.6	3.4	17.0	37.4‡
% Attentional	9.4	5.7	8.0	1.8
Validity variables				
Youth classifications of disorder				
% Emotional	20.9	10.2	18.4	5.1 +
% Behavioural	10.8	8.0	10.2	0.6
% Attentional	4.0	3.5	3.9	0.0
Current use of medications				
% Anxiety/ depression	14.4	0.0	8.8	27.7‡
% Behaviour	7.9	2.9	6.0	4.9 +
% ADHD	12.3	1.7	8.2	15.9‡
Mental-health related service contacts				
% School	30.3	17.1	25.2	9.86†
% Social services	35.0	8.6	24.8	40.24‡
% Health professionals	42.6	8.0	29.2	62.10†
Family psycho-social risk				
Maternal depression: mean (SD)[a]	18.6 (6.5)	3.4 (3.3)	0.0 (1.0)	n/a
Family dysfunction: mean (SD)[a]	20.8 (5.4)	18.7 (4.8)	0.0 (1.0)	n/a

ADHD: attention-deficit hyperactivity disorder; CCTB: Canadian Child Tax Benefit File; MINI-KID-P: Mini International Neuropsychiatric Interview for Children and Adolescents-parent version; OCHS-EBS-B: Ontario Child Health Study Emotional Behavioural Scales-Brief checklist; SD: standard deviation.

[a]Converted to z-scores for analysis; +<0.05, †<0.01, ‡<0.001.

The past empirical studies cited above exhibit one or more of the following limitations: relatively small sample sizes, the absence of reliability comparisons, the use of data "eyeballing" to make inferences, and measurement error compressing associations downwards. We have overcome these limitations by: combining independent samples to obtain a relatively large number of participants, comparing the test-retest reliability of the two instruments in the same time interval (1 to 2 weeks); using methods to correct for the attenuating effects of measurement error and implementing formal empirical tests of measurement equivalence.

Methods

The two studies used in this report are described in Table 1: study A[15] and study B.[16]

Participants

Of the 464 parent participants in the two studies, 452 had complete assessment data on the MINI-KID-P and OCHS-EBS-B on two occasions (Table 1)—meeting the inclusion criterion for our analyses. Net response in study A was 17% and in study B, 50%.

Ethical Considerations

Study A was approved by Research Ethics Committees at McMaster University, the School Boards and participating service providers. Informed written consent was obtained from participants. Study B procedures, including consent and confidentiality requirements, were approved by the Chief Statistician at Statistics Canada and were conducted according to the Statistics Act. Informed verbal consent was obtained.

Mental Health Disorders

Ontario Child Health Study Emotional Behavioural Scales-Brief Version. The OCHS-EBS-B is a 25-item problem checklist for completion by parents/caregivers of 4 to 17 years old. It is comprised of rating scales that measure three types of mental health problems[17]: emotional (8 items), behavioural (10 items), and attentional (7 items). Item response options are 0, 1, 2 ("never or not true," "sometimes or somewhat true," and "often or very true"). Raw scores are summed to form scale scores measuring each type of problem. The development and evaluation of the OCHS-EBS-B scales appear in the Appendix.

To convert the OCHS-EBS-B scale scores into classifications of disorder, their frequency distributions were examined in the 2014 Ontario Child Health Study[16] to find the score cut-points which identified children in the top 7.8% (emotional: 6+), 5.7% (behavioural: 6+), and 3.4% (attentional: 10+) that matched the world-wide prevalence of child major depressive or anxiety disorders, CD or ODD,

and ADHD.[18] These cut-points were applied in each study to convert the OCHS-EBS-B scale scores into classifications of disorder independent of the MINI-KID-P.

Mini International Neuropsychiatric Interview for Children and Adolescents. The MINI-KID-P (parent version) and MINI-KID-Y (youth version) are SDIs that assess DSM-IV-TR disorders in children aged 6 to 17 years with a reference period of 6 months, except for CD which is one year.[19] Respondent specific test-retest reliability based on kappa (κ[20]) goes from 0.67 (MDD) to 0.77 (ADHD) for parents and from 0.46 (ADHD) to 0.64 (MDD) for youth.[15] In our study, MINI-KID-P classifications of disorder were coded as (1) present or (0) absent and grouped as follows: emotional (MDD or GAD), behavioural (CD or ODD) and attentional (ADHD) to match the OCHS-EBS-B problem types.

Construct Validity Variables. Construct validity is a general approach to validation that works from a set of fundamental principles grounded in the philosophy of science to evaluate the meaning (informational content) of a measured construct by examining its strength of association with other variables.[21] We focus on variables expected to be positively associated with the disorders and test measurement equivalence by comparing associations between latent variable measures of disorder based on the MINI-KID-P and OCHS-EBS-B classifications of disorder with latent variable measures of youth reported classifications of disorder; and three related constructs measured as latent variables: current use of prescription medications for child mental health problems; child mental health-related service contacts with schools, service organizations or health professionals; and family psycho-social risk.

Youth Classifications of Disorder. Youth were administered the MINI-KID-Y at times 1 and 2 and classified with disorder as present or absent at each administration using the same groupings created for parents: emotional (MDD or GAD), behavioural (CD or ODD), and attentional (ADHD).

Current use of Prescription Medications for Mental Health Problems. At time 1, a latent variable measure of medication use was based on three binary indicator variables derived from parent interview responses to a stem question "Is <child> currently taking any prescribed medication?" and follow-up "What does <child> take this medication for: (a) hyperactivity, (b) behavioural problems, (c) depression or anxiety?"

Mental Health-related Service Contacts. At time 1, a latent variable measure of mental health-related service contacts was based on three binary indicator variables derived from parent interview responses to the following groups of questions: (1) "In the past 6 months, did <child>, or you see or talk to anyone from school about any emotional or behavioural problems?" (contact with schools); (2) "Has <child> ever been away overnight in: a foster or group home? a detention centre or juvenile centre; a police station or jail?", "In the past 6 months did <child>, or you see or talk to anyone: from the Children's Aid Society; from court, probation or other justice services; from some other service for children with emotional or behavioural problems?" (contact with service organizations); (3) "During the past 6 months … did <child>, or you, personally see any of the following about <child> emotional, behavioural or learning problems: a paediatrician or family doctor (general practitioner); a psychiatrist, psychologist or social worker?" (contact with health professionals).

Family Psycho-social Risk. At time 1, a latent variable measure of family psycho-social risk was based on scale score measures of maternal depressed mood and family dysfunction derived from parent responses on a self-completed questionnaire. In study A, maternal depressed mood is measured by the 12-item adaptation of the Center for Epidemiologic Studies Depression (CES-D) Scale[22] and in study B, by the K6.[23] Used often in empirical studies, the characteristics and psychometric properties of both measures are well described. In studies A and B, internal consistency reliabilities (ICR: alpha) are 0.89 (CES-D) and 0.79 (K6).

Family dysfunction is measured in both studies by the 12-item general functioning subscale of the McMaster Family Assessment Device.[24] The items describe family behaviour and relationships in six dimensions: problem-solving, communication, roles, affective responsiveness, affective involvement and behavioural control. Item responses go from (1) strongly agree to (4) strongly disagree. Positive items are reverse-coded; and all items, summed to form scale scores. ICRs were 0.89 (study A) and 0.83 (study B). The scale correlates predictably with alternative measures of family functioning.[25]

Analysis

We use the confidence interval (CI) approach to test our hypotheses of statistical equivalence in the reliability and validity of the MINI-KID-P and OCHS-EBS-B.[26] This test computes between-instrument mean differences in the parameters of interest (i.e., κ, β), the standard errors (SEs) of these differences and their 90% CIs. These 90% CIs are placed within an equivalence region called the smallest effect size of interest (SESOI) that represents between-instrument differences that are smaller than what is considered meaningful. If the $100(1-2\alpha)\%$ CI for the difference lies entirely within the equivalence region, the alternative hypothesis (H_1) of no difference is accepted, as illustrated here: $[H_0 \mid\leftarrow H_1 \rightarrow\mid H_0]$. We have set the SESOI for testing between-instrument differences in reliability and validity at ± 0.10 and ± 0.15, respectively. In our view, reliability differences within 0.10 are too small to be meaningful while validity differences within 0.15 represent <2.25% of explained variance associated with the disorders and are too small to be meaningful.

Test-Retest Reliability. We use κ, a chance-corrected measure of agreement scaled from 0 to 1 to quantify the test-retest reliabilities of the MINI-KID-P and OCHS-EBS-B for classifying the disorders**.** In estimating κ, we account for dependency attributable to the same respondent completing both instruments by using the weighted least-squares (WLS[27]) approach.

To determine if the reliability of the two instruments are statistically equivalent between studies, recruitment samples (general population vs. clinical), child sex or age, we extend the WLS approach in separate analyses to estimate and compare, one at a time, the sub-group specific reliabilities for the MINI-KID-P and OCHS-EBS-B (e.g., males vs. females). In these analyses, to increase statistical power by decreasing the SEs of the estimates, the three disorder types are analysed altogether rather than one at a time. Each of the 452 participants contributes three data points, one for each disorder type, to the reliability estimates. The number of participants remains the same ($n = 452$) but the number of observations triples ($442 \times 3 = 1{,}326$). The nesting or clustering of observations within participants (intra-class correlation) is taken into account in the statistical tests.

Agreement Between Instruments. We use κ to quantify between-instrument agreement in the classification of disorders at times 1 and 2. Based on the instrument test-retest reliabilities observed in our study and the sample correlation method described by Moss[28], we also show the estimated levels of agreement and their 95% CIs corrected for attenuation due to measurement error.

Construct Validity. We use β coefficients derived from structural equation models (SEMs) to estimate and compare the strength of association of the construct validity variables with the disorders classified by the MINI-KID-P and OCHS-EBS-B. All of the variables in the SEMs are measured as latent variables. A latent variable is unobserved—a prediction based on the pattern of associations among a set of observed indicator variables. The variance associated with a latent variable is free of measurement error in the sense that residual variance associated with the indicator variables, not predictive of the latent variable, is removed.

As illustrated in Figure 1A, each disorder (emotional, behavioural and attentional) classified by the MINI-KID-P and OCHS-EBS-B is measured as a latent variable based on its time 1 and 2 classifications (indicator variables). The latent variable measures of disorder based on youth interviews draw on their time 1 and 2 MINI-KID-Y classifications in exactly the same way. As a result, the SEMs involving youth classifications are specific to each disorder. The construct validity variables are also measured as latent variables (Figure 1B). These latent variables draw only on indicator variables assessed at time 1. In the comparison of the MINI-KID-P and OCHS-EBS-B, each disorder is modelled as a function of the same construct validity variable (e.g., current use of prescription medications examined separately for emotional, behavioural and attentional disorders). These three construct validity variables are expected to be associated generally with all of the disorders. Although prescription medications are linked with individual disorders, we created a latent variable measure of prescription medications to reduce the impact of misclassification errors associated with parent responses.

To test for statistical equivalence in the βs, we specify separate SEMs for each disorder. There are three latent variable measures in each SEM: two for the same disorders assessed by each instrument and one for the construct validity variable. We used MPlus 7.4[29] for the analyses. MPlus offers a generalized measurement component applicable to dichotomous and ordered categorical indicator variables.[30] Although the measurement scale is the same for both instruments (0, 1), the prevalences (variances) are different. As a result, we use standardized variables in the SEMs to ensure that the comparisons of βs are made on a commensurate scale. Adequate model fit was defined as values ≥0.98 for the comparative fit index (CFI, range 0 to 1.0) and ≤0.05 for the root mean squared error of approximation (RMSEA).

Results

The sample characteristics and distribution of variables appear in Table 1. In study A versus B, there are more 12 to 18 year olds, greater levels of socio-economic disadvantage (e.g., single parents) and a higher prevalence of child disorder.

Table 2 shows the test-retest reliability outcomes. Differences between the MINI-KID-P and OCHS-EBS-B ($\kappa_1 - \kappa_2$) in the classification of disorders goes from −0.064 (emotional disorder) to +0.053 (attentional disorder) with an average difference of −0.007 (not shown). Only behavioural disorder (−0.041) meets the criteria for statistical equivalence (see $_a$). In the subgroup analysis, the most extreme differences are associated with the recruitment sample: −0.072 (general population) and +0.025 (clinic). The average difference in the subgroup analyses, −0.027. The criteria for equivalence are met for Study A, males, youth aged 13 to 18 years and for both studies combined.

Table 3 shows agreement between the MINI-KID-P and OCHS-EBS-B on the classifications of disorder at times 1 and 2. Observed κ goes from 0.481 (attentional at time 1) to 0.721 (emotional at time 1). After correcting for attenuation due to measurement error, the corresponding estimates are substantially higher at 0.731 and 0.895, respectively.

Table 4 shows the βs associated with the MINI-KID-P and OCHS-EBS-B classifications of disorder when modelled as functions of the construct validity variables. Model fit (CFI) is ≥0.99 for all models (not shown). With one exception, the RMSEAs are ≤0.05 (Table 4). The average

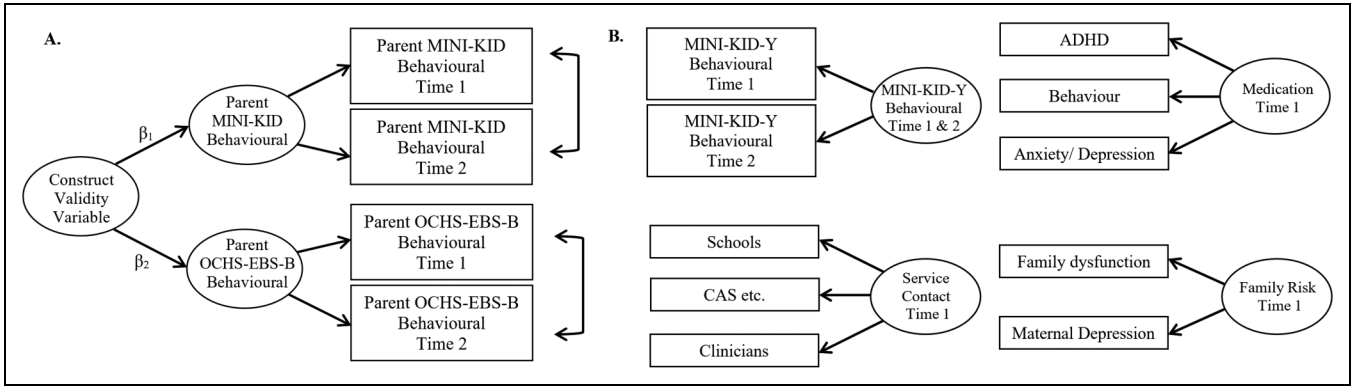

Figure 1. Illustration of latent variable measures developed for SEMs. (A) Parent assessment of behavioural problems based on MINI-KID-P classifications of CD or ODD at times 1 and 2 and OCHS-EBS-B classifications of behavioural problems at times 1 and 2. Construct validity tests if the 90%CI around $\beta_1 - \beta_2$ falls within −0.20 and+0.20. (B) The construct validity variables include youth assessments of behavioural problems based on MINI-KID-Y classifications of CD or ODD at times 1 and 2 (separate latent variables are developed for emotional and attentional problems); and three construct validity variables: current use of prescribed medications; service contacts with schools, service organizations or health professionals and family psycho-social risk. CI: confidence interval, MINI-KID-P: Mini International Neuropsychiatric Interview for Children and Adolescents interview-parent version; OCHS-EBS-B: Ontario Child Health Study Emotional Behavioural Scales-Brief checklist; SEM: structural equation models.

Table 2. Test-Retest Reliability (κ) of Parent MINI-KID-P and OCHS-EBS-B for Classifying Individual Child Disorders and Combined Disorders Grouped by Study, Recruitment Sample, Child Sex and Age.

Characteristics		MINI-KID-P κ_1 (SE)	OCHS-EBS-B κ_2 (SE)	$\Delta\kappa_1 - \kappa_2$ (SE)	Δ90% CI
Test-retest reliability of the disorders ($n = 452$)					
Disorder	Emotional	0.774 (0.039)	0.838 (0.031)	–0.064 (0.044)	–0.137 to 0.009
	Behavioural	0.774 (0.040)	0.785 (0.040)	–0.011 (0.050)	–0.094 to 0.071$_a$
	Attentional	0.685 (0.068)	0.632 (0.078)	+0.053 (0.093)	–0.100 to 0.206
Test-retest reliability in sub-groups ($n = 1356$)					
Study	Study A ($n = 831$)	0.771 (0.027)	0.795 (0.025)	–0.024 (0.034)	–0.080 to 0.032$_a$
	Study B ($n = 525$)	0.533 (0.105)	0.550 (0.121)	–0.017 (0.154)	–0.271 to 0.237
Recruitment sample	General ($n = 1074$)	0.675 (0.047)	0.747 (0.044)	–0.072 (0.057)	–0.166 to 0.022
	Clinic ($n = 282$)	0.734 (0.041)	0.709 (0.042)	+0.025 (0.055)	–0.066 to 0.116
Child sex	Female ($n = 729$)	0.739 (0.038)	0.798 (0.033)	–0.059 (0.046)	–0.135 to 0.017
	Male ($n = 627$)	0.786 (0.034)	0.788 (0.034)	–0.002 (0.042)	–0.071 to 0.067$_a$
Child age (years)	4–12 ($n = 489$)	0.756 (0.045)	0.787 (0.043)	–0.031 (0.052)	–0.117 to 0.055
	13–18 ($n = 867$)	0.766 (0.031)	0.796 (0.029)	–0.030 (0.039)	–0.094 to 0.034$_a$
Both studies ($n = 1356$)		0.763 (0.025)	0.793 (0.024)	–0.030 (0.031)	–0.081 to 0.021$_a$

Note: κ = kappa; $_a$90% CI within equivalence boundaries (−0.10 to +0.10); CI: confidence interval; MINI-KID-P: Mini International Neuropsychiatric Interview for Children and Adolescents-parent version; OCHS-EBS-B: Ontario Child Health Study Emotional Behavioural Scales-Brief checklist; SE: standard error.

difference between instruments (β_1–β_2) is −0.020 (not shown). The criteria for statistical equivalence are met in 5 of 12 comparisons.

Discussion

In our study, average between-instrument differences in test-retest reliability (Δκ) were small and favoured the OCHS-EBS-B over the MINI-KID-P. However, only 5 of 12 comparisons met the criteria for statistical equivalence. Correcting for attenuation due to measurement error revealed that between-instrument agreement in the classifications of disorder based on κ was >0.80 for emotional and behavioural disorders and slightly lower for attentional disorder. This very high level of the agreement indicates substantial overlap between instruments in their classifications of disorder—a finding that can only become apparent when adjustments are made for measurement error. Our analyses of construct validity differences between instruments mirrored the pattern of results observed for reliability: on average Δβ −0.020 favouring the OCHS-EBS-B over the MINI-KID-P. However, the criteria of statistical equivalence were met in only 5 of 12 comparisons. On balance, these findings are suggestive that the reliability and validity of the OCHS-EBS-B and MINI-KID-P completed by parents (mostly mothers) are similar if not equivalent for classifying disorders in children. However, there are several limitations and challenges associated with this work.

Table 3. Agreement (κ) Between the MINI-KID-P and Parent OCHS-EBS-B on the Classification of Child Disorders: Observed and Corrected for Attenuation due to Measurement Error ($n = 452$).

Disorder	κ (SE)	Corrected κ (95% CI)
Time 1		
Emotional	0.721 (0.041)	0.895 (0.805 to 0.978)
Behavioural	0.675 (0.047)	0.866 (0.762 to 0.962)
Attentional	0.481 (0.068)	0.731 (0.564 to 0.891)
Time 2		
Emotional	0.678 (0.043)	0.842 (0.742 to 0.932)
Behavioural	0.664 (0.048)	0.852 (0.745 to 0.949)
Attentional	0.521 (0.082)	0.792 (0.631 to 0.947)

CI: confidence interval; MINI-KID-P: Mini International Neuropsychiatric Interview for Children and Adolescents-parent version; OCHS-EBS-B: Ontario Child Health Study Emotional Behavioural Scales-Brief checklist; SE: standard error; κ: kappa.

Study Limitations and Challenges

One, nonresponse overall was particularly high in study A and we lack the information to evaluate the representativeness of this sample. In study B, the prevalence of OCHS-EBS-B disorders should be the same as the OCHS survey population but it is lower, suggesting that there was either selective nonresponse or the reliability sample was chosen in a geographical area at lower risk for the disorder. The lower estimates of reliability in study B are the product of low prevalence: κ estimates of reliability are compressed downwards under conditions of low prevalence, particularly <5.0%. Nonresponse could affect estimates overall or exert a differential effect on reliability and validity across instruments. It seems unlikely that nonresponse would have a differential effect on the instruments' reliability and validity—there is no obvious mechanism.

Two, we are unaware of any studies in psychiatry that have attempted to demonstrate psychometric equivalence

Table 4. Structural Equation Model Regression Results of Parent Classifications of Disorder (MINI-KID-P and Parent Reported OCHS-EBS-B) on Latent Variable Covariates.

Validity variables	Disorder	MINI-KID-P β_1 (SE)	OCHS-EBS-B β_2 (SE)	$\Delta(\beta_1 - \beta_2)$ (SE)	Δ90% CI
Youth disorders					
Attentional	Attentional	0.573 (0.117)	0.613 (0.120)	−0.040 (0.115)	−0.230 to 0.150
Behavioural	Behavioural	0.587 (0.085)	0.631 (0.083)	−0.044 (0.071)	−0.161 to 0.073[b]
Emotional	Emotional	0.560 (0.073)	0.508 (0.077)	+0.052 (0.057)	−0.042 to 0.146[b]
Current use of prescription medications	Attentional	0.854 (0.047)	0.898 (0.041)	−0.044 (0.054)	−0.133 to 0.045[b]
	Behavioural	0.669 (0.070)	0.670 (0.069)	−0.001 (0.063)	−0.105 to 0.103[b]
	Emotional[a]	0.775 (0.060)	0.886 (0.049)	−0.111 (0.046)	−0.187 to −0.035[b]
Mental health-related service contacts	Attentional	0.714 (0.069)	0.835 (0.053)	−0.121 (0.071)	−0.238 to −0.004
	Behavioural	0.847 (0.039)	0.806 (0.041)	+0.041 (0.036)	−0.018 to 0.100[b]
	Emotional	0.897 (0.032)	0.900 (0.031)	−0.003 (0.028)	−0.076 to 0.016[b]
Family psycho-social risk	Attentional	0.386 (0.096)	0.343 (0.101)	+0.043 (0.088)	−0.102 to 0.188[b]
	Behavioural	0.517 (0.072)	0.607 (0.066)	−0.090 (0.048)	−0.169 to −0.011[b]
	Emotional	0.503 (0.074)	0.422 (0.074)	+0.081 (0.059)	−0.016 to 0.178[b]

[a]Root mean square error of approximation = 0.053; [b]$\Delta\beta$ 90% CI within equivalence boundaries (−0.20 to +0.20); CI: confidence interval; MINI-KID-P: Mini International Neuropsychiatric Interview for Children and Adolescents-parent version; OCHS-EBS-B: Ontario Child Health Study Emotional Behavioural Scales-Brief checklist; SE: standard error.

between alternative instruments used to classify psychiatric disorder. Accordingly, there are no precedents for identifying the SESOI. The SESOIs selected in our study are conservative which has resulted in low statistical power. Doubling the sample size to 900 would make it possible for us to show that the 90% CI for observed between-instrument differences in $\kappa = \pm 0.04$ would fall within an equivalence boundary of ± 0.10. To date, there have been no measurement studies of this size in child psychiatry and the prospects for funding such studies are low.

Three, in testing for construct validity differences between the two instruments, we selected three variables that were "independent" of the specific questions/items and instrumentation approach of the MINI-KID-P and OCHS-EBS-B. We expected that these variables would exhibit relatively large, nonspecific associations with the three disorders classified by the instruments. Although the use of the MINI-KID-Y violates the principle of independence (i.e., relies on the same content and similar instrumentation as the MINI-KID-P which should positively bias its association with the MINI-KID-P), we believe that the absence of between-instrument differences in the βs associated with the MINI-KID-Y classification of emotional disorder represents evidence of equivalence.

Four, the absence of a strong, evidentiary-based consensus on the variables that might be used to quantify instrument differences in construct validity is a serious challenge.[4] To the best of our knowledge, no studies have compared the construct validity of alternative SDIs. As a result, we are not able to compare our findings with previous work and have a little context for interpreting the effect sizes observed in our study.

Five, in this article, we focus exclusively on the measurement objective of classifying child disorders based on the OCHS-EBS-B scale scores converted to binary ratings of disorder and compared with classifications of disorder derived from lay-administered SDIs. In contrast, clinicians may use SDIs as part of the diagnostic process to identify the disorder and its origin and to engage patients and formulate treatment plans. This is an entirely different context beyond the scope of our study.

Conclusions

In the years ahead, we believe that governments will focus on comprehensive approaches to addressing the mental health problems of children.[31] This will require the ability to monitor these problems in both the general population and in those accessing mental health services. A prerequisite for monitoring these problems will be simple, brief, inexpensive measurement instruments that are psychometrically sound, flexible in their administration, able to measure mental health problems as dimensional phenomena or, converted to binary measures, classify corresponding disorders as present or absent on a par with SDIs. Importantly, these instruments will need to exhibit near-identical psychometric properties when used in the general population and clinical samples. We believe that the OCHS-EBS-B go a long way to satisfying these extensive and detailed requirements.

Given the striking differences in cost and burden between SDIs and problem checklists, it is surprising how little research has been directed towards testing the extent to which problem checklists can substitute for SDIs in classifying disorder as present or absent. Additional studies addressing this question are urgently needed to provide clinicians, administrators and researchers with appropriate evidence for making cost-effective decisions about using problem checklists to monitor child mental health needs in the general population and in clinical settings.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Institute of Human Development, Child and Youth Health, research operating grant 125941 from the Canadian Institutes of Health Research (CIHR) and Health Services Research Grant 8-42298 from the Ontario Ministry of Health and Long-Term Care (MOHLTC) and with support from MOHLTC, the Ontario Ministry of Children and Youth Services and the Ontario Ministry of Education. Dr. Duncan is supported by a Research Early Career Award from Hamilton Health Sciences Foundation.

ORCID iD

Laura Duncan https://orcid.org/0000-0001-7120-6629

Supplemental Material

Supplemental material for this article is available online.

References

1. Pickles A, Angold A. Natural categories or fundamental dimensions: on carving nature at the joints and the rearticulation of psychopathology. Dev Psychopathol. 2003;15(3)529-551.
2. Coghill D, Sonuga-Barke EJS. Annual research review: categories versus dimensions in the classification and conceptualisation of child and adolescent mental disorders—implications of recent empirical study. J Child Psychol Psychiatry. 2012; 53(5):469-489.
3. Angold A, Erkanl A, Copeland W, et al. Psychiatric diagnostic interviews for children and adolescents: a comparative study. J Am Acad Child Adolesc Psychiatry. 2012;51(5):506-517.
4. Boyle MH, Duncan L, Georgiades K, et al. Classifying child and adolescent psychiatric disorder by problem checklists and standardized interviews. Int J Methods Psychiatr Res. 2017;26(4):e1544.
5. Drill R, Nakash O, DeFife JA, et al. Assessment of clinical information: comparison of the validity of a structured clinical interview (the SCID) and the clinical diagnostic interview. J Nerv Ment Dis. 2015;203(6):459.
6. Nordgaard J, Revsbech R, SæBYE D, et al. Assessing the diagnostic validity of a structured psychiatric interview in a first-admission hospital sample. World Psychiatry. 2012;11(3):181.
7. Thienemann M. Introducing a structured interview into a clinical setting. J Am Acad Child Adolesc Psychiatry. 2004;43(8): 1057-1060.
8. Groves RM. Nonresponse rates and nonresponse bias in household surveys. Public Opin Q. 2006;70(5):646-675.
9. Duncan L, Comeau J, Wang L, et al. Research review: test-retest reliability of standardized diagnostic interviews to assess child and adolescent psychiatric disorders: a systematic review and meta-analysis. J Child Psychol Psychiatry. 2019; 60(1):16-29.
10. Boyle MH, Duncan L, Georgiades K, et al. The 2014 Ontario Child Health Study Emotional Behavioural Scales (OCHS-EBS) part II: psychometric adequacy for categorical measurement of selected DSM-5 disorders. Can J Psychiatry. 2019;64(6):434-442.
11. Boyle MH, Offord DR, Racine Y, et al. Adequacy of interviews versus checklists for classifying childhood psychiatric disorder based on parent reports. Arch Gen Psychiatry. 1997;54(9):793-799.
12. Gould MS, Bird H, Jaramillo BS. Correspondence between statistically derived behavior problem syndromes and child psychiatric diagnoses in a community sample. J Abnorm Child Psychol. 1993;21(3):287-313.
13. Jensen PS, Watanabe HK, Richters JE, et al., Scales, diagnosis and child psychopathology: iI comparing the CBCL and the DISC against external validators. J Abnorm Child Psychol. 1996;24(2):151-168.
14. Jensen PS, Watanabe HK. Sherlock Holmes and child psychopathology assessment approaches: the case of the false-positive. J Am Acad Child Psychiatry. 1999;38(2):138-146.
15. Duncan L, Georgiades K, Wang L, et al. Psychometric evaluation of the Mini International Neuropsychiatric Interview for Children and Adolescents (MINI-KID). Psychol Assess. 2018;30(7):916-928.
16. Boyle MH, Georgiades K, Duncan L, et al. The 2014 Ontario Child Health Study—methodology. Can J Psychiatry. 2019; 64(4):237-245.
17. Ogundele MO. Behavioural and emotional disorders in childhood: a brief overview for paediatricians. World J Clin Pediatr. 2018;7(1):9-26.
18. Polanczyk GV, Salum GA, Sugaya LS, et al. Annual research review: a meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. J Child Psychol Psychiatry. 2015;56(3):345-365.
19. Sheehan DV, Sheehan KH, Shytle RD, et al. Reliability and validity of the Mini International Neuropsychiatric Interview for Children and Adolescents (MINI-KID). J Clin Psychiatry. 2010;71(3):313-326.
20. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46.
21. Cronbach LJ, Meehl P. Construct validity in psychological tests. Psychol Bull. 1955;52(4):281-302.
22. Poulin C, Hand D, Boudreau B. Validity of a 12-item version of the CES-D [centre for epidemiological studies depression scale] used in the National Longitudinal Study of Children and Youth. Chronic Diseases in Can. 2005;26(2–3):65-72.
23. Kessler RC, Barker PR, Colpe LJ, et al. Screening for serious mental illness in the general population. Arch Gen Psychiatry. 2003;60(2):184-189.
24. Epstein N, Baldwin L, Bishop D, et al. The McMaster Family Assessment Device. J Marital Fam Ther. 1983;9(2):171-180.
25. Georgiades K, Boyle MH, Jenkins JM, et al. A multilevel analysis of whole family functioning using the McMaster Family Assessment Device. J Fam Psychol. 2008;22(3):344-354.

26. Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: a tutorial. Adv Methods Pract Psycholog Sci. 2018;1(2):259-269.
27. Barnhart HX, Williamson JM. Weighted least-squares approach for comparing correlated kappa. Biometrics. 2002;58(4):1012-1019.
28. Moss J. (2019). Correcting for attenuation due to measurement error. arXiv preprint arXiv:1911.01576. 2019 Nov 2 Accessed June 18, 2021 from https://arxiv.org/abs/1911.01576
29. Muthén LK, Muthén BO. Mplus user's guide. 6th ed. Los Angeles (CA): Muthén & Muthén; 2016.
30. Muthén BO. A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. Psychometrika. 1984;49:115-132.
31. Boyle MH, Duncan L, Georgiades K, et al. Tracking children's mental health in the 21st century: lessons from the 2014 OCHS. Can J Psychiatry. 2019;64(4):232-236.