Chapter 3

Validation of the Multiple Language Versions of the Hopkins Symptom Checklist-37 for Refugee Adolescents

Abstract

The objective of this study is to provide preliminary psychometric properties of the Hopkins Symptom Checklist-37 (HSCL-37A) for refugee adolescents. The HSCL-37A is a modification of the well-known HSCL-25 and assesses symptoms of internalizing and externalizing problems that have been associated with reactions to trauma. Four independent heterogeneous samples (N = 3890) of unaccompanied refugee minors, immigrants and native Dutch and Belgian adolescents were assessed at school. The confirmative factor analyses, per language version, support the two-factor structure of internalizing and externalizing behavior. The total and subscales show good internal consistency and acceptable test-retest reliability in spite of the heterogeneous sample populations. The construct, content and criterion validity of the HSCL-37A were also examined and found to be good. The findings of this study suggest that the HSCL-37A is a reliable and valid instrument to be used among culturally diverse refugee adolescents to assess emotional distress and maladaptive behaviors.

Introduction

During the last 20 years, there has been a substantial influx of immigrants and refugees migrating to Europe (Eurostat, 2002). This has led to more and more schools in Dutch and Belgian metropolitan areas that provide education for children and adolescents who do not

fluently speak the language of the host country. The transition from one country to another implies changes and difficulties such as the loss of social networks, changes in work status as well as encountering discrimination what can be very distressing (Vinokurov, Trickett, & Birman, 2002).

The traditional higher - order latent structure of internalizing (over controlling) and externalizing (under controlling) problems have for many years been a useful framework for emotional distress and maladaptive behaviors of children and adolescents (e.g., Achenbach & Edelbrock, 1978; Southam-Gerow & Kendall, 2002). In recent years, Krueger and colleagues (2001) confirmed the usefulness of this dichotomy in explaining the covariance among adult mental health and personality disorders. Moreover, Miller and colleagues (2003) have proceeded to put forward an internalizing/externalizing model to explain the reactions of traumatic stress among adult combat veterans. The internalizing/externalizing model seems to give an adequate framework in which traumatic stress reactions and/or (comorbid) psychopathology can be understood.

The literature on the mental health of refugee adolescent provides a depiction of high prevalence rates of psychosocial symptoms reported by refugee adolescents (Felsman, Leong, Johnson, & Felsman, 1990; Sack et al., 1993; Sourander, 1998; Smith, Perrin, Yule, Hacam, & Stuvland, 2002). The most frequently reported symptoms are somatic complaints, anxiety, depression, and (post)traumatic stress reactions. Unaccompanied refugee minors (URM) run an especially high risk for developing psychopathology due to separation from primary caregivers, exposure to sequential stressful events, limited educational opportunities, and conditions in asylum centers during a very vulnerable developmental period (Felsman et al., 1990; Sourander, 1998). High comorbidity has been documented between reactions to traumatic stress and other disorders such as depression (Sack et al., 1993) and anxiety (Warshaw et al., 1993). Significant adults in the lives of adolescents (i.e., caregivers, teachers) often report a lower prevalence of internalizing problems than the adolescents themselves having difficulty determining to what extent the adolescents suffer from psychological distress.

On the other hand, perceiving the disturbing nature of externalizing problems is not difficult. Adolescents with conduct problems have been found to be referred much sooner/often to professional mental healthcare services than adolescents with internalizing problems (Wu et al., 1999). The literature concerning conduct problems of refugee adolescents is very limited. Allwood, Bell-Dolan, & Husain (2002) found a strong association between witnessing of organized violence and exhibiting aggressive behavior. Jensen and Shaw (1993) suggest that adolescents who have witnessed or taken part in a war are more likely to show delinquent or anti-social behavior. This opinion is, however, not supported in four studies which evaluated the delinquent and aggressive behaviors of refugee adolescents (Raboteg-Saric, Zuzul, & Kerestes, 1994; Mollica et al., 1997; Rousseau, Drapeau, & Corin, 1998; Sourander, 1998). Different authors (i.e., Pynoos & Nader, 1993) report that adolescents may temporarily show increased risk behavior following the witnessing/experiencing of a traumatic event. Moreover, several studies have found high levels of comorbidity between externalizing behavior and experiencing traumatic stress reactions among American adolescents (Deykin & Buka, 1997; Wozniak et al., 1999)

The “pathway” to professional mental healthcare for refugee adolescents has more barriers than for native adolescents in host countries (e.g., Howard & Hodes, 2000). There is sufficient evidence in the literature suggesting that young people in general that are in need of psychological support or treatment do not receive it (i.e., US Department of Health and Human Services, 1999; Cuffe, Waller, Cuccaro, Pumareiga, & Garrison, 1995) or only when the symptoms have progressed and are perceived by significant adults in their lives (e.g., Wu et al., 1999). The psychological suffering of Unaccompanied Refugee Minors (URM) can go completely unnoticed due to the absence of parents or permanent caregivers, language difficulties and living in minimally adult supervised residential settings.

Mental healthcare professionals in host countries are often hindered in acquiring accurate information concerning the mental health status of refugee adolescents due, in part, to language difficulties, no medical/psychological background information and to a lack of translated reliable and valid diagnostic instruments. Approaching refugee youth with long psychological questionnaires/interviews can be very overwhelming (Barenbaum, Ruchkin, & Schwab-Stone, 2004). Brief, translated psychological instruments that measures, reliably and

validly, the broader reactions associated with the traumatic stress that refugee adolescents have experienced could be of great assistance to mental healthcare professional in the process of screening, diagnosing and monitoring the mental health status of this specific high-risk population.

There is a limited number of diagnostic instruments that can be used with immigrant adolescents to measure psychosocial distress and maladaptive behaviors such as the Youth Self Report (Achenbach, 1991a) and Strengths and Difficulties questionnaire (Goodman, 1997). These two instruments have been used with refugee adolescents from specific countries to measures emotional and behavior problems (Fazel & Stein, 2002; Mollica et al., 1997) However, as far as known by the authors, these checklists (as well as others used with refugee children (i.e., Smith et al., 2002) have not been validated for culturally diverse adolescent populations following the five dimensions of equivalence for cross-cultural validation of an instrument proposed by Flahtery et al. (1988).

An increasing number of studies have been conducted with refugee culturally homogeneous samples (Papageorgiu et al., 2000; Thabet & Vostanis, 1999) or samples from two different countries of origin (Rousseau & Drapeau, 1998). Smith and colleagues (2003) have validated the Revised-Impact of Events Scale with older children from Bosnia. However, the same instrument yielded less reliable results with older children and adolescents from Rwanda (Dyregrov, Gupya, Gjestad, & Mukanoheli, 2000) which clearly illustrates that when a measure has been validated for one immigrant or refugee population, it does not implicitly infer that the measure is valid and reliable for all refugee and immigrant populations.

Because no validated questionnaire was available, modifications were made to one of the well-known instruments that has been used with refugees/non-western populations of adults over the last 15 years, the Hopkins-Symptom Checklist-25 (Lie, 2002; Mollica et al., 1987). The objective of this study was to provide preliminary information concerning the psychometric properties of the modified version of the HSCL-25 (Winokur, Winokur, Rickels, & Cox, 1984), the HSCL-37 for adolescents. Twelve externalizing items have been added to the 25 original items to measure problematic behavior in adolescents, which may be trauma-related.

Methods

Context of the validation study

In the years preceding 2001, there was a dramatic increase in the number of URM living in the Netherlands, peaking at 15,000 in 2001. Because there was (and still is) a lack of research studies on the mental health and service utilization of URM, a national and longitudinal research project "Unaccompanied Refugee Minors and Dutch Mental Healthcare Services" was started among unaccompanied refugee minors living in the Netherlands and their guardians, teachers and professional mental healthcare providers. A secondary aim of the project was to validate and standardize screening instruments for this specific population group. It was also possible to administer the HSCL-37A in an independent research project conducted by the Department of Orthopedagogics, Ghent University, Belgium that examined whether being unaccompanied is a risk factor for refugee children and adolescents to develop emotional and behavioral problems.

Samples

Dutch URM sample (n = 920). A national, longitudinal study was carried out with URM living in the Netherlands. Approximately 4000 URM were randomly selected from the Central Registrar of Nidos. Information about the study and permission waivers (available in translated versions) were sent to the guardians to discuss with the URM. Both the minor and his/her guardian gave written permission for the URM to participate. Roughly 2300 URM permission waivers were returned; 57% wished to participate, 15% refused, 12% did not participate for a wide range of practical reasons, 9% were transferred to a different residential setting, and 7% turned out to be untraceable. A total of 920 URM were present for participation. The final sample was statistically representative (data not shown) in all of the main characteristics (age, gender, country of origin type of residential setting) of the total URM population aged 12 to 18 year old in 2002 in the Netherlands. The URM came from 48

countries, predominantly Angola (43%), Sierra Leone (10%), and China (8%). Two-thirds of the sample had lived in the Netherlands for a period of 18 months or less. 45% of the URM sample had received 5 years or less of formal education in their country of origin. A follow-up (63% of the original sample participated) was conducted one year after the first assessment. An interview regarding mental healthcare was individually administered. At least three research assistants administered the questionnaires during one hour to groups of 10 URM.

Dutch normative sample (*n* = 1059*)*. Pupils from ten secondary and three tertiary trade schools throughout the Netherlands (schools had also taken part in the URM study) participated and functioned as a control group for the URM sample. Two weeks prior to administration of the instruments, informed consent letters were sent to the parents and adolescents asking for the voluntary and anonymous participation (27 students abstained from participation). The assessment of the Dutch sample took approximately 15 minutes.

Belgian immigrant /refugee adolescents sample (*n* = 1294). A large scale study was carried out with non-Dutch speaking immigrant adolescents in Flanders (Belgium) during November 2002 to May 2003. The adolescents came from 111 countries, predominantly Morocco (14%), Ghana (11%), and Turkey (9%). All schools received standard informed consent letters (translated versions were available) asking parents and students for voluntary and anonymous participation. In 2002, there were 42 secondary schools in Flanders which provided education for recently immigrated adolescents. Thirty-four schools were randomly chosen to participate in the study of which none declined. 65% of the number of recently (less than 1 year) immigrated adolescents (immigrants and refugees) in Flanders between 13-18 years of age, participated in the study. Only 1 student abstained from participation which was present on the day of assessment. There was a continuous stream of new students during the year, which render it very difficult to test the entire population. No attempt was made to test adolescents that were not present on assessment day. The assessment took place (1 hour) under supervision of two research assistants.

Belgian normative sample (*n* = 617). A control group of Belgian adolescents participated between January, 2003 and May, 2003 for the Belgium immigrant/refugee study. From the six Flemish provinces, 17 secondary schools were randomly selected to participate in the study. All schools received standard informed consent letters asking parents and students for voluntary and anonymous participation. To assemble a well-balanced normative sample of the Flanders adolescent population, the same percentage of Belgian adolescents and Immigrant/Refugee (I/R) adolescents per province took part in the study. In this way, there would not be an overrepresentation of Belgian adolescents living in urban or rural areas. Furthermore, the proportions for the different age and gender groups of the Belgian adolescents were carefully matched with those of the I/R sample so that the two groups were similar on these variables. Finally, per province the secondary schools that were chosen had students that were following all three educational track levels (trade, occupational and preparatory for university). No Flemish student refrained from participating.

Table 1.
Summary of Sample Characteristics

		Gender**		Age in years			Group**			Type of caregiver**	
	N	Boys	Girls	Mean	SD	Range	Natives	Immigrants Refugees	URM	Parental	Other
Total sample	3890	59.3	40.7	15.72	1.74	8-26	40.7	30.7	28.6	70	30
Dutch URM	920	72.8	27.2	15.68	1.49	8-20	0	0	100	0	100
Belgian immigrant/refugee adolescents	1294	53.9	46.1	15.41	1.88	10-26	1	89.1	9.9	84.3	15.7
Dutch adolescents	1059	56.8	43.2	15.72	1.54	13-21	90.1	9.9	0	97.3	2.7
Belgian adolescents	617	54.6	45.4	16.46	1.92	13-21	97.9	2.1	0	97.6	2.4

Note. ** percentages

Measures

The HSCL-37A (Bean, Eurelings-Bontekoe, Derluyn, & Spinhoven, 2004a) was modified to render the instrument multi-cultural and adolescent friendly. A 4-point rating scale in literal terms (*not/never* = 1, *sometimes* = 2, *often* = 3, *always* = 4) was used to indicate the severity of symptoms, feelings or behaviors. The literal terms of the Likert scale was improved by placing different colored balls increasing in size above the literal rating scale to clarify "quantity" of feelings. Secondly, items were (if needed) simplified to adapt the questionnaire to the (Dutch) language abilities of this population based on a vocabulary lists developed for immigrant adolescents to the Netherlands, and thirdly, the questionnaires were translated and presented in a bilingual form (Dutch-foreign language). It was necessary to have the questionnaires in bilingual form because many of the refugee adolescents had limited written knowledge of their own language and learned the Dutch language quickly allowing them to use both languages to be able to better comprehend the item.

The HSCL-37A, SLE and RATS questionnaires were translated into the most prevalent languages of URM in the Netherlands: Albanian, Amharic, Arabic, Badini, Chinese, Dari, Dutch, English, Farsi, French, German, Mongolian, Portuguese, Russian, Servo-Croatian, Soerani, Somali, Spanish and Turkish. All written forward translations were done by professionally employed translators. Every translation was controlled for grammatical and idiomatic errors on two different occasions by two different translators. The translated questionnaires were reviewed orally with professional interpreters who where regularly involved in treatment sessions of traumatized adult refugees to control the quality of the translations, to ensure that the original meaning was conveyed in the items, and to attempt to achieve semantic equivalence of the HSCL-37A. No written back-translations were done in this study. Instead an oral item-by-item analysis took place with trained interpreters from mental health services. All of the instruments were tested in a pilot study. If an adolescent filled in a bilingual version of the instruments, the bilingual version of the questionnaire was recorded. If an adolescent completed the Dutch version only, Dutch was recorded as language of the questionnaire(s).

The internalizing scale of the HSCL-37A can be divided into ten anxiety questions (items 1, 2, 5, 9, 12, 16, 19, 22, 26, 29) and fifteen depression questions (items 6, 10, 13, 15, 17, 20, 23, 24, 27, 30, 31, 32, 33, 35, 36). The scale for externalizing behavior (items 3, 7, 11, 14, 18, 21, 34, 4, 8, 25, 28, 37; min = 12, max = 48) can be used to attain a total score for externalizing behavior. The externalzing items *bullies*, *steal things*, *intentionally hurting someone*, *starts fights*, *destroying others property* correspond with five criteria from the diagnosis for a Conduct Disorder. The item *easily angered* and *argues often* correspond with two criteria from the diagnosis for an Oppositional Defiant Disorder according to the DSM-IV (American Psychiatric Association, 1994). The other five items are related to substance abuse (*use of alcohol in the weekend*, *use of alcohol through the week*, *smoking cigarettes*, *use of sedatives*, and *use of drugs*). The total score of the HSCL-37A consists of all of the 37 items (min. = 37, max. = 148). Percentile scores and severity classifications are available in the user's manual (Bean et al., 2004a).

The *Stressful Live Events questionnaire* (SLE) (Bean, Derluyn, Eurelings-Bontekoe, Broekart & Spinhoven, in press; Bean et al., 2004b) consists of 12 dichotomous (yes/no) questions and an open question on the occurrence of stressful life events of relevance for adolescent refugee minors (e.g., "Have you ever experienced a war or an armed military conflict going on around you in your country of birth?" or "Has someone ever hit, kicked, shot at or some other way tried to physically hurt you?"). Experiencing a traumatic event is the first criterion of the A cluster of the DSM-IV for PTSD (APA, 1994). The overall average total score of 6.5 of the SLE has been validated in 5 independent studies (Bean et al., 2004b).

The *Reactions of Adolescents to Traumatic Stress* (RATS) (Bean et al., in press; Bean et al., 2004c) is a self-report questionnaire developed to assess posttraumatic stress reactions defined in the DSM-IV (APA, 1994) with culturally diverse adolescents. The twenty-two item scale can be divided into three subscales: intrusion (six items), avoidance (nine items) and hyper-arousal (seven items) which correspond to the 17 criteria in a PTSD diagnosis. Internal reliability for the URM sample for the total score, intrusion, numbing/avoidance and hyperarousal was correspondingly, .88, .85, .69, and .73. Twelve-month test-retest reliability was for total score .61 ($p < .001$). Using a confirmatory factor analysis, the three-factor

structure was verified in the URM sample with a loss of only 3% of the explained variance. Similar results were found confirmed in the other 3 samples.

The self-report version for 11- to 16-year olds of the *Strengths and Difficulties Questionnaire* (SDQ) (Goodman, 1997) is a screening questionnaire that measures twenty-five attributes divided into five subscales: emotional symptoms, conduct problems, inattention-hyperactivity, peer problems, and pro-social behavior. Research shows that the SDQ has an acceptable reliability and validity (Goodman, 2001). The SDQ was also available in the languages of the immigrant/refugee adolescents in Belgium. In this study, the internal reliability (Cronbach's alpha) of the total score of the multiple language versions of the SDQ ranged from .62 - .79, with an average value of .63. Average sub-scales reliability was low-to-unacceptable .68, emotion symptoms .42, peer problems for the total population.

Indicators of Psychopathology

The criteria "referral" and "utilization of MHC" have been documented as being important in the evaluation of psychopathology in children and adolescents (i.e., Verhulst & Van der Ende, 1997). For this reason, (a) self-reported need for mental healthcare (MHC); (b) need for professional MHC for the URM; evaluated by the legal guardian; (c) need for professional MHC for the URM; evaluated by the teacher; (d) self-reported utilization of MHC by URM; and (e) referral to MHC services by a legal guardian were utilized as external criteria of psychopathology. The URM were individually interviewed in Dutch about their needs and mental health use. They were also able to read the questions in one of the language that have been mentioned above. Guardians and teachers received short questionnaires on need for professional MHC and referral to MHC services by URM which they filled-in and returned by mail.

A strong, significant, and positive relationship should exist between the HSCL-37A total and the SDQ total scores because these two scales measure the same construct. There should also be a strong association between the HSCL-37A internalizing scale and the RATS because as reported earlier high co-morbidity has been found between PTSD on the one hand and general anxiety/depression on the other. The correlation between the externalizing score (measuring trauma-associated acting out behaviour) of the HSCL-37A and the RATS scores should be present but weak. The total SLE score should be positively related to the total score of the HSCL-37A and subscales, since trauma is related to psychopathology (Allwood et al., 2002; Tiet et al.,1998).

Procedures

Ethical approval for both Belgian studies was given by the Ethics Committee of the Faculty of Psychology and Educational Sciences, Ghent University and by the Medical Ethics Committee of the Leiden University Medical Center, Leiden University to conduct the Dutch URM study.

Testing of the Belgian and Dutch normative samples took place in small groups (10-25 young people) during school time. The URM were assessed at schools, if possible. Approximately 20% of the URM were not tested at schools. URM were also assessed (in groups of 10) at the regional offices of Nidos, reception centers for refugees, and residential settings. Demographic information on the URM in the Netherlands was supplied by the Nidos Foundation (legal guardian of all of the URM living in the Netherlands). The rest took part anonymously and answered written questions that provided demographic characteristics about themselves.

Data Analysis

Descriptive statistics were used to give summary descriptions of the socio-demographic characteristics of the sample. Confirmatory factor analyses, per language version, were calculated using the Multiple Group Method (MGM) procedure of the Simultaneous Components Analysis (SCA) (Kiers, 1990) to verify the factorial validity of the HSCL-37A (all cases with missing data were removed). MGM is closely related to the rotation of component weights to perfect congruence and the cross-validation of components weights (Ten Berge, 1986). SCA is based on the *same* set of weights for the variables in all populations enabling conclusions on the common components found across the samples. It is not a formal statistical test, such as the Maximum Likelihood estimation method. However,

this is not a serious objection because the null hypothesis of a factor model based on a small number of factors is invariably false as has been known since Browne (1969, p. 385). Failure to reject it merely means that the sample size has been too small (see McCrae, Zonderman, Costa, Bond, & Paunonen, 1996 for a discussion). Internal consistency of the total scale and subscales of the HSCL-37A was calculated with Cronbach's α. Test-retest reliability was calculated for a twelve month interval for the URM sample only (n = 519). Pearson's product-moment correlations (two-tailed) were used to study the association between total and subscale scores of the HSCL-37A and the scores on the remaining questionnaires. Differences between groups were determined by using ANOVA's and effect sizes Cohen's d (Cohen, 1988). A maximum of ten percent of missing items was allowed to still be able to extrapolate the total or subscale scores of all scales.

Results

Factorial Validity

The factor structure of the HSCL-37A was tested with the Simultaneous Components Analysis (SCA). The scale consisting of internalizing items was established based on the results of a previous factor analysis on a large item pool and opinions of several experienced clinicians (Derogatis et al., 1974). The externalizing items made up the second scale. For the total sample, a principal component analysis (PCA) was used with Varimax rotation (oblique) to simple structure which allowed for correlation between the two factors (Kiers, 1990) which yielded a model that explained 33.1% of the total variance. The SCA-MGM analysis based on the two a priori factors showed that the multiple group components explained 32.7% of the variance, implying a small acceptable discrepancy of only .4%.

Separate MGM analyses were conducted on the Portuguese, French, Chinese, English, Arabic, Dutch, and Russian language versions. The amount of variance that was lost in enforcing the a priori factor structure in comparison to the results of an explorative PCA in the separate language versions was very limited, ranging from 2.2 % in the Chinese version to .4% in the Dutch version. Due to the limited number of completed questionnaires ($n < 100$) in Badini, Servo-Croatian, Albanese, Turkish, Soerani, Dari, Farsi, Amharic, Somali, and Mongolian, no individual MGM's could be conducted for these languages. The two-factor model is confirmed in all the separate MGM analyses per language (Table 2).

Internal consistency

The internal consistency (Cronbach's alpha) of the HSCL-37A indicates a high degree of homogeneity among items comprising the total and subscales in the separate language versions. The internal consistency of the total scale of the HSCL-37A in the total sample was .90 and of the individual language versions ranged from .95 to .84. This is an exceptionally high alpha, despite the high degree of heterogeneity in the samples. The alpha's for the subscales and apart language versions can be found in Table 2.

Temporal Stability

The test-retest scores are utilised to provide an indication of scale stability and consistency over time. The coefficients show that the HSCL-37A scales are reasonably stable ($r > .50$) over time in measuring internalizing and externalizing behavior (Table 2) not deviating from findings of other studies with the same time interval (see Cheng & Nicholas, 1998 for a discussion).

Content validity

Content validity is a measure of the relevance of the items with regard to that behavior which they aim to measure. The HSCL-37A claims to measure internalizing (anxiety and depression symptoms) and externalizing behavior. The choice of items to measure anxiety and depression was based on the expertise of clinicians with experience in the treatment of patients with anxiety and depression (Derogatis et al., 1974). All items of the HSCL-37A correspond with the DSM-IV criteria for anxiety, depression, and behavior symptoms. The 12 externalizing items correspond with the five criteria of conduct disorder and the two criteria of the oppositional-defiant disorder, as defined in the DSM-IV (APA, 1994). The content validity of the HSCL-37A is good.

Table 2.

Summary of Confirmatory Factor Analyses and Reliability Analyses per language version

	Two factor				Total scale				internalizing					externalizing			
Language	*n*	*EV*	*LV*	α	r_{ii}	r_{it}	r_{ab*}	*n*	α	r_{ii}	r_{it}	r_{ab*}	*n*	α	r_{ii}	r_{it}	r_{ab*}
Total sample	3019	33.1%	.4%	.90	.18	.05-.59	.63	3126	.92	.30	.36-.66	.64	3524	.75	.20	.13-.45	.53
Dutch	1640	31.4%	.4%	.88	.17	.13-.55		1670	.90	.27	.30-.63		1771	.75	.22	.15-.50	
Portuguese	326	29.5%	.6%	.90	.18	.01-.62		342	.91	.28	.22-.64		374	.62	.12	.10-.42	
English	215	32.6%	.8%	.91	.20	.17-.63		230	.91	.29	.19-.68		298	.72	.18	.14-.51	
French	163	33.0%	.5%	.91	.21	.05-.63		166	.91	.30	.34-.68		220	.71	.20	.24-.57	
Arabic	127	34.9%	1.3%	.92	.22	-.04-.70		141	.92	.31	.23-.75		184	.67	.16	.16-.50	
Turkish	118	NA	NA	.92	.24	.00-.70		123	.92	.33	.33-.70		151	.66	.16	.15-.50	
Russian	111	43.6%	.7%	.95	.29	.01-.80		119	.95	.46	.41-.79		137	.58	.12	.11-.42	
Chinese	95	47%	2.2%	.92	.24	.13-.69		96	.93	.37	.43-.71		106	.74	.24	.24-.59	
Spanish	47	NA	NA	.84	.13	-.04-.62		47	.78	.12	.05-.51		53	.76	.25	.21-.69	
Farsi		NA	NA	NA	NA	NA		44	.89	.25	.17-.67		NA	NA	NA	NA	
Albanese		NA	NA	NA	NA	NA		29	.88	.23	.07-.75		NA	NA	NA	NA	
Servo-Croatian		NA	NA	NA	NA	NA		22	.93	.33	-.14-.77		NA	NA	NA	NA	
Dari	23	NA	NA	.91	.21	.10-.74		24	.90	.28	.11-.68		30	.67	.17	.07-.61	
Amharic	18	NA	NA	.91	.20	-.04-.70		21	.92	.32	.37-.70		26	.61	.19	.02-.73	
Somali		NA	NA	NA	NA	NA		16	.94	.37	-.03-.85		NA	NA	NA	NA	
German		NA	NA	NA	NA	NA		17	.91	.30	.09-.77		NA	NA	NA	NA	
Mongolian	11	NA	NA	.86	.14	-.34-.90		11	.86	.17	-.34.89		12	.60	.18	-.28-.76	

Note. EV = Explained Variance with PCA; LV = Loss of Explained Variance with MGM; α = Alpha coefficient; r_{ii} = Mean inter-item correlation; r_{it} = Range item-total correlations; r_{ab*}=Test- re-test reliability calculated for 12 month interval for URM sample only, *n* =519; NA = not able to analyze, more than one item with zero variance.

Table 3.
Intermeasure correlations

	Total RATS		Total SLE		Total SDQ	
URM	*(n)*	r	*(n)*	r	*(n)*	r
Total HSCL-37A	(771)	.74	(819)	.39		
internalizing	(761)	.79	(812)	.41		
externalizing	(780)	.32	(835)	.12		
Dutch natives						
Total HSCL-37A	(1058)	.75	(1057)	.48		
internalizing	(1058)	.76	(1057)	.36		
externalizing	(1058)	.23	(1057)	.39		
Belgian immigrants/refugees						
Total HSCL-37A	(870)	.66	(1167)	.38	(1117)	.65
internalizing	(854)	.68	(1149)	.38	(1101)	.64
externalizing	(886)	.33	(1192)	.22	(1141)	.43
Belgian natives						
Total HSCL-37A	(596)	.67	(615)	.38	(612)	.70
internalizing	(596)	.67	(614)	.30	(611)	.64
externalizing	(597)	.31	(616)	.34	(613)	.42

Note. All correlations are significant at the .001 level and two-tailed.

Construct validity

Construct validity is a measure of the relationship between the instrument and variables that, on theoretical grounds, are expected to correlate with the measured variable. In construct validation, three processes are used to establish construct validity; (1) convergent validity: high correlations between a particular scale and others that in theory measure the same construct, (2) discriminant validity: low associations between the scale under study and other measures that should theoretically not be related, and (3) factorial validity: supports the theory-based grouping of items when a particular construct is complex. Table 3 shows the intercorrelations (two-tailed) between the HSCL-37A total and subscale scores, the RATS total score and the SLE total score for the URM sample and native Dutch sample. In Table 3, the intercorrelations are presented between the SDQ total score (the SDQ was only administered in the Belgian studies), the RATS total score and the SLE total score for the immigrant/refugee sample and native Belgian sample.

As hypothesized, the HSCL-37A total scores and internalizing scale scores show significant and positive correlations with the RATS total scores, SLE total scale scores and SDQ total scores. The significant and positive relationship between the externalizing scale scores and the other scale scores is weaker, but still present. The relationship between the total, internalizing, and externalizing scores on the HSCL-37A and the total number of experienced events on the SLE is significant and positive. These findings are applicable to all samples.

The mean scores of girls are expected to be significantly higher than that of boys. Girls reported significantly higher internalizing ($F(1,3646) = 74.96$, $p < .001$, $d = .29$) and externalizing mean scores ($F(1,3718) = 25.03$, $p < .001$, $d = .17$) than boys. There are contradictory findings in the literature concerning age and emotional distress. Age, in this study, seemed to play a small role with respect to total mean scores, with older adolescents ($\leq$ 17 years) scoring significantly higher than younger ($\geq$ 14, 15 years) for internalizing ($F(3,3597) = 30.50$, $p < .001$, $d = .39$-$.23$) and externalizing problems ($F(3,3668) = 16.96$, $p < .001$, $d = .31$-$.10$). Because of the numerous risk factors overshadowing the lives of URM, it was expected that URM would score significantly higher than immigrant/refugee and native adolescents living with at least one parent. This expectation is partly confirmed. URM reported significantly higher internalizing mean scores ($F(2,3661) = 269.24$, $p < .001$, $d = .87$-$.78$) on the HSCL-37A than the immigrants/refugees and natives, but significantly lower externalizing mean scores than native adolescents ($F(2,3733) = 273.37$, $p < .001$, $d = .72$) .

Table 4.
External criteria influencing HSCL-37A internalizing and externalizing scores

	internalizing						externalizing				
	n	*M*	*SD*	*t (df)*	*p*	*d*	*n*	*M*	*SD*	*t (df)*	*p*	*d*
URM: Need for MHC												
Need for MHC	438	54.33	12.75	8.84 (592)	<.001	.83	448	15.63	3.22	1.92 (605)	ns	.18
No need for MHC	156	44.03	11.79			.	159	15.06	3.09			
Guardian: Need for MHC												
Need for MHC	95	57.47	12.07	6.03 (492)	<.001	.69	97	16.52	3.69	3.38 (506)	<.001	.38
No need for MHC	399	49.05	12.28				411	15.28	3.13			
Teacher: Need for MHC												
Need for MHC	115	54.63	14.22	4.06 (401)	<.001	.45	120	15.96	2.93	3.17 (413)	<.01	.34
No need for MHC	288	48.62	13.09				295	14.94	2.99			
URM: MHC Utilization												
Utilization of MHC	96	54.79	14.14	2.68 (682)	<.01	.30	95	15.97	3.09	1.40 (697)	ns	.16
No utilization of MHC	588	50.98	12.70				604	15.47	3.22			
Referral: MHC by guardian												
Referred to MHC	59	59.28	13.54	5.73 (496)	<.001	.80	59	16.56	2.97	2.79 (511)	<.01	.39
Not referred MHC	439	49.53	12.09				454	15.36	3.14			

Criterion validity

Criterion validity refers to the association between the instrument with some form of external or outside criterion that is supposed to measure the same construct. Criterion validity can be further divided into two types; predictive and concurrent. Only the concurrent validity of the HSCL-37A was addressed in this study. Five indicators of psychopathology were utilized as external criteria (see Indicators of Psychopathology description in Questionnaires section). The results shown in Table 4 suggest that the HSCL-37A discriminated well, consistently, and significantly between URM that report having a need for psychosocial help and URM that did not report having a need for psychosocial help.

Discussion

The results indicate that the HSCL-37A is a psychometrically sound screening measure of internalizing and externalizing problems experienced by a heterogeneous population of refugee adolescents. The data has been collected among four independent sample stretched across the Netherlands and Flemish Belgium. The psychometric properties of the HSCL-37A demonstrate invariance of factor structure in a heterogeneous sample, strong reliability, and good validity which is remarkable considering the diversity of the populations.

The layout modifications (bilingual and visual/literal rating scale) of the instrument made the HSCL-37A comprehensible for adolescents from a variety of cultures. In a small number of individual cases, lengthy explanations of the meaning/nuances of the items were necessary, especially with “almost” illiterate adolescents (1 per group of 25 refugees/immigrants). It is not clear if errors in understanding the questions might not be visible in the data. Only the Spanish version of the HSCL-37A had obvious less internal consistency on all subscales. This could be due to the fact that the translation was in European Spanish and adolescents came from South American countries which speak a different dialect of Spanish. European Spanish was used because of the wide differences in dialects in all Spanish speaking countries (the same holds for American English and European French which did not show lower reliability levels).

The two factors showed strong reliability and good validity considering the diversity of the sample populations (for example; adolescent from 35 different counties filled in the French version, adolescents from 57 countries filled in the English version and adolescents from 20 different countries filled in the Arabic version). The preliminary validity findings suggest that the HSCL-37A can discriminate consistently and significantly between refugee adolescents that do need to utilize MHC services and those who do not.

The brevity of the HSCL-37A takes into account the importance of not overburdening apprehensive adolescents and allows for quick, repeated measurements to assist with determining initial and enduring refugee adolescent symptomatology. When the HSCL-37A is used in juncture with the SLE and RATS questionnaires a preliminary assessment of the global mental health of refugee adolescents can be reliably and validly assessed. In all settings, one must be aware that the instrument may trigger emotional distress. Therefore, adequate crisis and/or follow-up MHC should be arranged prior to administration to protect the integrity of the adolescents. The HSCL-37A is not meant to be used alone as a diagnostic instrument for internalizing distress or behavioral problems. Clinical observations and additional assessment are important in establishing a valid diagnosis and making treatment recommendations.

Methodological challenges

There were several methodological challenges of this study. Written back-translations of the language versions were not done, deviating from standard protocol which can be seen as a limitation of the study. Back-translation is the method that is used to verify semantic equivalence of translated measures (see Mallinckrodt & Wang, 2004 for a discussion). However, a back-translation does not implicitly guarantee that the content equivalence of the translated instrument has been established (Flahtery et al., 1988). A great amount of effort in this study was spent on ensuring the content equivalence of the items of the HSCL-37A for different cultures.

The number of instruments that were used were limited to a minimum for a number of reasons; (a) short attention spans of the refugee adolescents, (b) the amount of time needed to

explain and administer the three instruments took around 15 minutes of the testing time, (c) the substantial amount of time and effort used by the refugee adolescents to complete only three questionnaires, and (d) the ethical issues related to the administration of long instruments with severely traumatized individuals which might induce emotional distress. Additional measures would have enhanced the quality of the study and would have been useful in determining the divergent validity of the HSCL-37A which will need to be evaluated in future studies.

The stability (test-retest) of the HSCL-37A was calculated over a longer interval (12 months) than the usual 8-week interval resulting in a lower temporal stability then is desired. However, it could be expected after one year that many changes (due to developmental changes, stressful life events, transfers, change in residential status, and therapeutic interventions) would have taken place in the constantly changing lives of URM which could have led to even lower stability levels.

Because no standardized diagnostic interview was utilized in this study, the sensitivity and specificity of the HSCL-37A could not be evaluated. Preferably, a standardized diagnostic interview is used in combination with questionnaires to determine the presence and severity of psychopathology. However, referral of children and adolescents to psychiatric services has been used as a "golden standard" instead of a diagnostic interview (e.g., Nolan et al., 1996). It was not feasible in the URM study to administer a diagnostic interview for the reasons that have been listed above and that there is no validated psychiatric diagnostic interview available in all of the languages of (refugee) the adolescents who took part in this study. Furthermore, the use of diagnostic interviews invokes itself a host of methodological issues such as classifying culture-specific disorders and ensuring "the semantic and psycholinguistic equivalence of psychiatric symptoms across cultures" (Cheng, 2001). Even so, the preliminary validity findings suggest that the HSCL-37A is able to discriminate between adolescents that do and do not need to utilize mental health services.

Self-report questionnaires such as the HSCL-37A yield less diagnostic information than extensive structured interviews and therefore should be used only to indicate clinically elevated levels of internalizing and externalizing problems and not to diagnosis anxiety, depression or conduct disorder. Additional information should be collected regarding the mental health of the adolescent from the viewpoint of significant adults (caregivers/teachers) in the environment of the adolescent. This information is crucial in assessing the degree of impairment in daily functioning and the severity of the symptoms of adolescents.