International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 23(2): 161–183 (2014)
Published online 15 April 2014 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1426

Clinical use of the Kessler psychological distress scales with culturally diverse groups

YVONNE STOLK,[1] IDA KAPLAN[2] & JOSEF SZWARC[3]

1 Research Consultant, Victorian Foundation for Survivors of Torture, Brunswick, Vic., Australia
2 Direct Services Programme, Victorian Foundation for Survivors of Torture, Brunswick, Vic., Australia
3 Research and Policy Programme, Victorian Foundation for Survivors of Torture, Brunswick, Vic., Australia

Key words
K6, K10, psychological distress, screening scale, cultural equivalence, primary care, mental health

Correspondence
Josef Szwarc, Manager, Policy and Research, Victorian Foundation for Survivors of Torture, 4 Gardiner Street, Brunswick, Vic., 3056, Australia.
Telephone (+61) 410 529 217
Email: szwarcj@foundationhouse.org.au

Received 8 November 2012; revised 15 February 2013; accepted 25 February 2013

Abstract

The Kessler 10 (K10) and embedded Kessler 6 (K6) was developed to screen for non-specific psychological distress and serious mental illness in mental health surveys of English-speaking populations, but has been adopted in Western and non-Western countries as a screening and outcome measure in primary care and mental health settings. This review examines whether the original K6/K10's validity for culturally diverse populations was established, and whether the cultural equivalence, and sensitivity to change of translated or culturally adapted K6/K10s, has been demonstrated with culturally diverse client groups. Evidence for the original K6/K10's validity for culturally diverse populations is limited. Questions about the conceptual and linguistic equivalence of translated/adapted K6/K10s arise from reports of changes in item connotation and differential item functioning. Evidence for structural equivalence is inconsistent, as is support for criterion equivalence, with the majority of studies compromising on accuracy in case prediction. Research demonstrating sensitivity to change with culturally diverse groups is lacking. Inconsistent evidence for the K6/K10's cultural appropriateness in clinical settings, and a lack of clinical norms for either majority or culturally diverse groups, indicate the importance of further research into the psychological distress construct with culturally diverse clients, and the need for caution in interpreting K6/K10 scores. *Copyright © 2014 John Wiley & Sons, Ltd.*

Introduction

The Kessler 10 (K10) and embedded Kessler 6 (K6) are increasingly used as screening and outcome measures in primary care (Hickie *et al.*, 2002; Carra *et al.*, 2011) and mental health settings (Sakurai *et al.*, 2011; O'Connor *et al.*, 2012; Sunderland *et al.*, 2012a), although originally developed to screen for non-specific psychological distress and serious mental illness (SMI) in English-speaking (ES) population surveys (Kessler *et al.*, 2002; Furukawa *et al.*, 2003). However, the population of ES countries has become increasingly culturally diverse, with acknowledgement of Indigenous peoples, and growth in migrant and refugee resettlement (UNESCO World Report, 2009). It is important therefore to examine the K6 and K10's (hereafter abbreviated as K6/K10) applicability to culturally diverse groups in clinical settings. The term "culturally diverse" will be used as a collective term for indigenous, migrant, refugee, ethnic, racial, and linguistic groups (UNESCO World Report, 2009), except when a specific group is discussed.

The need for accurate mental health screening of people of culturally diverse backgrounds is demonstrated by research that shows disparities in access to mental health services (MHSs), prevalence of mental disorders, delays in treatment seeking, and involuntary admissions (Office of the Surgeon General, 2001; Morgan *et al.*, 2004; Stolk *et al.*, 2008; Vos *et al.*, 2009; Kim *et al.*, 2011; Gone and Trimble, 2012). Moreover, people of refugee background may have elevated risks of mental disorder resulting from pre-arrival trauma, forced migration, personal loss, time in detention centres, and other post-arrival stressors (Davidson *et al.*, 2008; Kirmayer *et al.*, 2011). As some culturally diverse groups are more likely to seek help from primary care services for mental health problems (Mereish *et al.*, 2012), culturally appropriate screening in these settings may aid in earlier detection and treatment of mental disorders, thereby reducing suffering and burden (Kirmayer *et al.*, 2011).

This paper therefore aims to investigate the cultural appropriateness of the K6/K10's use in clinical settings with culturally diverse groups. We examine: (a) whether validity and reliability for culturally diverse populations was established during development of the original K6/K10; (b) the validity, reliability, and cultural equivalence of translated or culturally adapted K6/K10s in clinical settings; and (c) the K6/K10's sensitivity to change with culturally diverse groups.

Method

Medline, PsycINFO, and Academic Search Complete databases from 2000 to 2012 were searched for peer-reviewed, English-language publications, using combinations of the search terms Kessler, K6, K-6, K10, K-10, psychological distress, and Kessler scale development. To identify publications relating to culturally diverse groups the following terms were added: rac*, ethnic, minorit*, indigenous, cultur*, divers*, translat* language*, refugee, asylum, primary care, mental health outcomes, sensitivity to change and intervention. Searches were also conducted on the Internet, of K6/K10 studies on the United States (US) National Comorbidity Survey (NCS, 2012) website, and of reference lists in included papers. The final studies selected were of adults aged 18 years and over, used translated or culturally adapted K10s, and provided evidence relating to validity, reliability, cultural equivalence, and/or sensitivity to change. Studies were excluded that provided no source, citation or explanation of translation or cultural adaptation procedures, as were studies of culturally diverse groups that did not report English proficiency inclusion or exclusion criteria, or use of translated instruments, as cultural equivalence could not be evaluated. Both clinical and epidemiological studies were included as both provided pertinent psychometric and cultural information.

Validity, reliability and cultural equivalence criteria

This review evaluates the cultural equivalence of translated or culturally adapted K6/K10s to ensure that the original scales' validity and reliability were maintained. Translation and cultural adaptation can change an instrument's validity and reliability, precluding cross-cultural comparisons of test scores. A translated instrument may be viewed as a new measure, requiring validation in the same way as the original (Van de Vijver, 1998; American Educational Research Association (AERA) *et al.*, 1999). To facilitate evaluation of translated and adapted instrument, we have aligned types of validity and cultural equivalence in Table 1, with definitions and methods for achieving these properties. Construct validity is defined as primary, and is interrelated with, and supported by evidence from test content, response processes, internal structure and relations to other variables (AERA *et al.*, 1999). Cultural equivalence may be broadly defined as "the extent to which constructs hold similar meanings … across cultural groups" (Arnold and Matus, 2000, p. 122). However, multiple definitions of cultural equivalence exist (Johnson, 1998), and Table 1 shows key typologies we have extracted from recent literature that may be subsumed under cultural equivalence, with brief definitions and methodologies. Culturally equivalent instruments are expected to demonstrate construct equivalence, which encompasses: conceptual and linguistic equivalence; structural, item and scalar equivalence, criterion or predictive equivalence, method and administrative equivalence, and normative equivalence. Ideally, evidence for more than one type of equivalence is provided (Van de Vijver and Leung, 2011). Types of cultural equivalence and methodologies may show some overlap. Reliability refers to a scale's internal consistency, shown by Cronbach's alpha (Arnold and Matus, 2000).

Results

Initial searches for Kessler scales yielded 1052 papers, from which five key studies on the K6/K10's development and cultural sensitivity were identified. Addition of search terms for culturally diverse groups reduced the number

Table 1. Relationship between, and methods for demonstrating, test validity and cultural equivalence

Validity[a,d,f]	Definition	Method	Types of cultural equivalence	Definition	Method
Construct validity – supported by evidence from: ↓	How well the instrument measures the concept it was designed to measure		Construct	The construct under investigation has the same meaning, and relevance in the target as in the source culture	
– *Test content*	How well items sample the universe of interest	Expert review of item relevance and representativeness	*Conceptual/ linguistic*	Translated items have the same meaning, familiarity and frequency of occurrence in both languages	Independent forward and back-translation; bilingual expert panel review; field testing; cognitive interview of respondents[b,c,e,g]
– *Response processes*	How respondents respond to tests	Cognitive interview of respondents, who are asked to 'think aloud'	*Item*	Items are equally likely to be endorsed; there is no differential item bias due to differences in cultural connotations, stimulus familiarity, social desirability or response styles	Cognitive interview; differential item analysis (DIF)[b,d,g,]
– *Internal structure*	Interrelations between test sub-components	Cronbach α; factor analysis	*Structural/ scalar*	There is identity of underlying dimensions. Total scale scores are comparable across cultures	Cronbach α, factor analysis to confirm identity of dimensions. Investigate mean group differences for response styles, DIF[d,g]
– *Relations to other variables*	Associations with other tests of the same construct	Criterion, or predictive validity studies	*Criterion/ predictive/ clinical*	The translated instrument discriminates cases from non-cases as accurately as the original	Conduct comparative predictive validity studies [b,f]
Requirements:					
Standardized administration	Administration methods, scoring and interpretation are specified in test manuals	Test manuals are developed	*Method/ administrative*	Administration methods have the same meaning, and are not differentially influenced by cultural and	Manuals for administration, scoring, and score interpretation are provided; administrators are trained;

(*Continues*)

Table 1. (Continued)

Validity[a,d,f]	Definition	Method	Types of cultural equivalence	Definition	Method
				linguistic factors. Scoring methods and score interpretation are culturally comparable	bilingual staff or interpreters are available[a,g]
Populations for which the test is appropriate are specified	Norms allow comparisons between test-takers of similar background	Norms are established	*Normative*	Norms are available for the target population.	Establish norms. Lacking norms, interpret scores with caution[a,d,e]

[a]AERA *et al.* (1999); [b]Arnold and Matus (2000); [c]Harkness *et al.* (2008); [d]Huysamen (2002); [e]Okawa (2008); [f]Sireci and Parker (2006); [g]Van de Vijver and Poortinga (2005).

of papers to 352. A focus on clinical validation and explanation of translation or cultural adaptation procedures resulted in a total of 21 studies that used translated K6/K10s. In addition, three studies considered cultural adaptations to the K6 or the K10, and two investigated sensitivity to change. Cultural adaptations involved consultation with Indigenous groups and modifications to ensure research protocols and the scale were culturally appropriate and used familiar language (Australian Institute of Health and Welfare (AIHW), 2009; Browne *et al.*, 2010; Wells *et al.*, 2006). It should be noted that this review's requirement that studies specify or cite translation/adaptation procedures, or English proficiency inclusion criteria, excluded a number of US studies examining race-ethnicity (e.g. Prochaska *et al.*, 2012).

Table 2 first summarizes the reports on the original K6/K10's psychometric properties, for comparison with the properties and translation procedures of the 24 studies using translated or culturally adapted K6/K10s (hereafter abbreviated as translated/adapted). Eleven studies are categorized as clinical and 11 as epidemiological. Two other studies that involved convenience (Grzywacz *et al.*, 2009) and chain referral samples (Sulaiman-Hill and Thompson, 2010), are not listed in Table 2 as they provided no psychometric validity data but contributed information on conceptual and linguistic equivalence (discussed later). To set the context for evaluation of translated/adapted K10s, the development of the K6/K10 is briefly reviewed to examine how validity was established for epidemiological and clinical settings, and whether cultural diversity was considered.

Validity and cultural sensitivity of the original K6/K10

The aim in developing the K6/K10 was to provide a screening scale sufficiently brief and sensitive to include in the US National Health Interview Survey (NHIS), and to identify the 10% of the population with severe psychological distress "in the clinical range" (Kessler *et al.*, 2002, p. 961). Content validity was ensured by selecting items from a wide array of existing non-specific psychological distress scales, and review by an expert panel. Because of the prevalence of items relating to anxiety and depression, items were sorted into those domains (Kessler *et al.*, 2002).

Reduced scales were pilot-tested in mail and telephone surveys, during which "race-ethnicity" was addressed by oversampling people with Hispanic surnames, and areas with a high proportion of "Blacks" (Kessler *et al.*, 2002). Factor analysis and item response theory were used to select items that loaded onto a single factor, and showed consistent severity values across socio-demographic

Table 2. Psychometric properties of the original and translated/culturally adapted K6 and K10s in clinical and epidemiological settings (K10 values shown in bold typeface)

Authors/ country/ groups/*N*	Languages	Translation protocol	Clinical criterion	Internal consistency	Cutoff/maximum score	Validity				
				α		Sensitivity	Specificity	PPV	AUC (95% CI)	Factor analysis
Reference studies										
Kessler *et al.* (2002). US NHIS Hispanic, Black groups oversampled. Final K6/K10: *N* = 1000	English	NA	SCID DSM-IV ADD, (GAF score ≤70)	K6: 0.89 **K10: 0.93**	Advised against cutoff scores. For surveys base severity score ranges on reference population				0.879 **0.876**	PFA 1 factor non-specific psychological distress
Kessler *et al.* (2003). US NHIS *N* = 155	English	NA	SMI: a SCID DSM-IV disorder (GAF score ≤60)	K6: 0.89 **K10: 0.93**	≥13/24	0.36	0.96		0.865 **0.854**	
Andrews and Slade (2001); and Furukawa *et al.* (2003) Australian NSMHWB *N* = 10,641	English	NA	CIDI DSM-IV ADD	K6: NR **K10: NR**	 **≥17/50**	 **0.81**	 **0.83**		0.89 **0.90**	
Aldworth *et al.* (2010) US NSDUH Total *N* = 45,000, sub-sample *n* = 1500 – White (W) 69% – Hispanic (H) 14% – Black (B) 11% – Other (O) 6%	NR	NR	SMI: SCID DSM-IV disorder (GAF score <50)	K6: NR	K6 alone: ≥17 ≥17 /24 + 8-item WHODAS ≥4	All: 0.387 All: 0.56 W: 0.491 H: 0.293 B: 0.804 O: 0.686	0.971 0.976 0.969 0.996 0.982 0.988		0.679 0.741 0.730 0.644 0.893 0.837	
Clinical studies										
Arnaud *et al.* (2010) France Emergency Department patients with alcohol-related disorders K6 *n* = 29; K10 *n* = 42	French	NR. From Canadian Community Health Survey	Alcohol Dependence/ Abuse module of MINI	K6: 0.76 **K10: 0.84**	≥10/SR NR **≥14/SR NR**	0.92 **0.95**	0.62 **0.54**	66.7 **65.5**	0.87 **0.77**	EFA: K6 2 factors **K10: 3 factors**
Baggaley *et al.* (2007). Burkina Faso Post-natal women *N* = 61	W. African French, Mooré, Dioula	WHO protocol	Psychiatrist interview ICD-10 depression	K6: 0.78 **K10: 0.87**	10/24 **≥12/40** **≥14/40**	59% **74%** **59%**	85% **76%** **91%**		0.75 **0.77**	

(Continues)

Table 2. Continued

Authors/ country/ groups/*N*	Languages	Translation protocol	Clinical criterion	Internal consistency	Cutoff/maximum score	Validity				
				α		Sensitivity	Specificity	PPV	AUC (95% CI)	Factor analysis
Carra *et al.* (2011) Italy Primary care. *N* = 605, sub-sample *n* = 147	Italian	WHO protocol	SCID DSM-IV ADD	K6: 0.84	≥7/24	0.82	70%		0.82 (0.75–0.89)	
				K10: 0.90	**≥12/40**	**87.2%**	**69.6%**		**0.83 (0.76–0.90)**	
					≥13/40	**79.5%**	**75.4%**		**0.77 (0.71–0.84)**	
Carra *et al.* (2011) continued		WHO protocol	SCID DSM-IV SMI	K6: 0.84	≥7/24	87%	82%		0.89(0.84–0.95)	
				K10: 0.90	**≥13/40**	**83.5%**	**87.1%**		**0.91(0.85–0.96)**	
Chowdhary and Patel (2010) India Care programme for HIV infected people, *N* = 109	Konkani, Marathi, Hindi	Stepwise protocol (Patel *et al.* 2008)	Psychiatrist-diagnosed ICD-10 CMD	K6: NR	**≥13/30**	68.4%	73.2%	58.7	0.78	
Donker *et al.* (2010a) Netherlands Patients of 65 general practitioners., *N* = 23,750, subsample *n* = 1607	Dutch	WHO protocol	CIDI-2.1 DSM-IV ADD	**K10: 0.94**	**≥20/50**	**0.80**	**0.81**	**0.63**	**0.87(0.869–0.873)**	
Donker *et al.* (2010b) Netherlands Internet-recruited respondents. *N* = 502, subsample *n* = 157	Dutch	WHO protocol	CIDI-2.1 DSM-IV depression	**K10; 0.90**	**≥29/50**	**0.81**	**0.67**		**0.81 (0.73–0.88)**	
Fernandes *et al.* (2011) India Rural pre-natal clinic. *N* = 194	Kannada	WHO WMHS guidelines	MINI DSM-IV depression	**K10: NR**	**≥6/40**	**1.00**	**0.81**	**0.48**	**0.95 (0.92–0.98)**	
Patel *et al.* (2008) India Primary care. *N* = 598	Konkani, & unspecified Indian languages	Stepwise protocol	CIS-R ICD-10 CMD	K6: 0.74	3–4/6	58%	91%	56%	0.845	
				K10: 0.82	**5–6/10**	**65%**	**89%**	**53%**	**0.877**	
					6–7/10	**54%**	**93%**	**62%**		
Sakurai *et al.* (2011) Japan Community members *n* = 147, psychiatric outpatients *n* = 17	Japanese	WHO protocol	Psychiatrist-diagnosed DSM-IV ADD	K6: 0.85	4–5/24	100.0%	68.7%		0.93 (0.88–0.98)	
				K10: 0.91	**9–10/40**	**100.0%**	**73.5%**		**0.94 (0.90–0.98)**	

Spies *et al.* (2009) South Africa Pre-natal clinic, *N* = 129	Afrikaans	Forward & back-translation	SCID DSM-IV Major depression	**K10: NR**	**≥21.5/SR NR**	**0.73**	**0.54**		**0.66**	
Tesfaye *et al.* (2010) Ethiopia Post-natal women *N* = 105	Amharic	Equivalent to WHO protocol	Psychiatrist interview CPRS CMD	K6: 0.86	4–5/SR NR	84.2%	82.7%		0.86 (0.76–0.97)	
				K10: 0.90	**6–7/SR NR**	**84.2%**	**77.8%**		**0.87 (0.78–0.97)**	
Epidemiological studies										
AIHW (2009); and Cunningham and Paradies (2012). Australia 2004–2005 National Aboriginal and Torres Strait Islander Health Survey *N* = 5757	English or Indigenous language (not specified)	Translation by indigenous facilitators as required	Self-reported mental illness (MI; & other stressors)		K5 SR 5–25.	K5 score	% with self-reported MI			
						Low: 5–7.9		6.5		
						Moderate 8–11.9		14.0		
						High 12–14.9		17.9		
						Very high 15–25		32.8		
Andersen *et al.* (2011) South Africa (WMHS) Total sample: *N* = 4077 – Black 76% – Other: White, Coloured, Indian/Asian 24%	Afrikaans, Zulu, Xhosa, N. Sotho, Tswana	Expert consensus	CIDI 3.0 DSM-IV ADD	K6: 0.48	≥10/30	70.2%	67.9%	20.7%	Total: 0.72 Black: 0.70 Other: 0.77	
				K10: 0.84	**≥16/50**	**70%**	**66.7%**	**23%**	**Total: 0.73** **Black: 0.71**	
					≥42/50	**4%**	**Specificity NR**	**50%**	**Other: 0.78**	
Browne *et al.* (2010) and Wells *et al.* (2006) New Zealand (WMHS) *N* = 7435 Maori *n* = 2595 Pacific people, *n* = 2236	English	Some interviewers assisted in other languages	CIDI 3.0 DSM-IV ADD CIDI 3.0 DSM-IV SMI	**K10: NR**	**NR/40**				**ADD**: **0.74 (0.72–0.76)** **SMI**: **0.80 (0.77–0.84)**	
Fassaert *et al.* (2009a) Netherlands, Amsterdam health survey,	Dutch, Turkish, Moroccan	Forward/ back translation	CIDI 2.1 DSM-IV ADD	**K10: 0.93**	**Dutch: ≥16.5/50**	**0.792**	**0.768**	**22.4**	**0.85 (0.79–0.92)**	**EFA, CFA[4] of combined data**:
					Turkish: ≥22.5	**0.795**	**0.748**	**46.8**	**0.80 (0.73–0.88)**	

(Continues)

Table 2. Continued

Authors/ country/ groups/*N*	Languages	Translation protocol	Clinical criterion	Internal consistency	Cutoff/maximum score	Validity				
				α		Sensitivity	Specificity	PPV	AUC (95% CI)	Factor analysis
Dutch *n* = 321 Turkish *n* = 213 Moroccan *n* = 191					**Moroccan: ≥22.5**	**0.824**	**0.776**	**26.4**	**0.88 (0.80–0.95)**	**1 factor**
Furukawa *et al*. (2008) Japan (WMHS) Health Survey *N* = 915	Japanese	WHO protocol	WMH CIDI DSM-IV ADD	K6: NR **K10: NR**	9–13/24 SSLR = 16 (95% *CI* = 6.1–34) **15–19/40 SSLR = 11 (95% *CI* = 2.3–32)**				0.94 (0.88–0.99) **0.94(0.88–0.995)**	
Ito *et al*. (2012) Japan Survey of complicated grief *N* = 915	Japanese	WHO protocol (Furukawa *et al*., 2008)		K6: 0.75						EFA BGQ and K6 items, CFA 1 K6 factor
Kessler *et al*. (2010) WMHS 14 countries Sample range: Lebanon 1031 – US 5692; Total *N* = 41770	Translations by each country	WHO WMHS guidelines	WHO CIDI DSM-IV ADD & SMI	K6: NR	Differing logistic regression formulae accurately predicted probability of SMI in different countries				Range: 0.76–0.89	EFA: 1 factor in all countries
Laube (2010) Australia 2002-2006 NSW health surveys. English *n* = 49365 Arabic *n* = 111 Chinese *n* = 275 Greek *n* = 116 Italian *n* = 112 Vietnamese *n* = 125	English, Arabic, Chinese, Vietnamese, Greek, Italian	Expert consensus		**K10: NR**						**EFA, parallel analysis, CFA: inconsistent patterns across groups**
Lee *et al*. (2012) Hong Kong Mental health survey Total *N* = 3014 sub-sample *n* = 153	Cantonese-Chinese	WHO protocol	SCID DSM-IV SMI	K6: 0.843	12–13/24	43.9%	94.6	36.2	0.69	EFA: 2 factors, depression & anxiety

Mitchell and Beals (2011) US American Indians *N*=3084	English	No cultural adaptation required	CIDI DSM-IV any mood disorder	K6: 0.83	≥13/24	29%	96%	0.77 (0.74–0.80)	CFA 1 factor
Silove *et al*. (2008) Timor Leste Displaced population *N*=1245	Indonesian, Tetun	Expert consensus	SCID DSM-IV diagnoses	**K10: NR**	**≥30/SR NR**	**82% of SCID non-psychotic disorders screen positive on K10 *and/or* HTQ**			

Note. Blank cells denote information not provided, instrument not used, or analysis not conducted. ADD = anxiety or mood disorders, AUC = area under receiver operating characteristic curve, BGQ = Brief Grief Questionnaire, CFA = confirmatory factor analysis, CIDI = Composite International Diagnostic Interview, CIS-R = Clinical Interview Schedule – Revised, CMD = common mental disorders, EFA = exploratory factor analysis, GAF = Global Assessment of Functioning, ICD-10 = International Classification of Diseases – 10th Revision, K6 = 6-item Kessler psychological distress scale, K10 = 10-item Kessler psychological distress scale, MINI = Mini-International Neuropsychiatric Interview, NA = not applicable, NHIS = National Health Interview Survey, NR = not reported, NSDUH = National Survey on Drug Use and Health, NSMHWB = National Survey of Mental Health and Well-Being, PFA = principal factor analysis; SMI = serious mental illness, SCID = Structured Clinical Interview for DSM-IV, SR NR = score range not reported, SSLR = stratum specific likelihood ratio, WHO = World Health Organization, WHODAS = WHO Disability Assessment Schedule, WMHS = participating study in the World Mental Health Survey.
$^{*}p < 0.05$; $^{**}p < 0.0001$.

factors, including race-ethnicity, so that scores had "the same meaning in all major segments of society" (Kessler *et al*., 2002, p. 965). Data were provided on symptom severity values for age, sex and education, but not race-ethnicity. Other US ethnic groups apparently were not sampled, and English proficiency inclusion criteria were not reported.

A clinical reappraisal survey of a small sub-sample of 155 respondents, selected from 1000 screened individuals, showed that the K6 and K10 accurately predicted cases and non-cases of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) anxiety and depression disorders (ADDs) on a structured diagnostic instrument, the Structured Clinical Interview for DSM-IV (SCID, First *et al*., 2002), and with a Global Assessment of Functioning (GAF, Endicott *et al*., 1976) score of ≤ 70 (Kessler *et al*., 2002). The K6 more efficiently predicted serious mental illness (SMI; defined as a DSM-IV disorder and a GAF score ≤60) than the K10 (Kessler *et al*., 2003; see Table 2).

The K6 was externally validated in the 1997 and 1998 NHIS and the K10 in the "nationally representative" (Kessler *et al*., 2002, p. 962) 1997 Australian National Survey of Mental Health and Well-being (NSMHWB). In the NSMHWB the K10 accurately predicted DSM-IV ADD (Andrews and Slade, 2001), based on the Composite International Diagnostic Interview (CIDI; Robins *et al*., 1988). However, non-English speakers (NES) were excluded (Henderson *et al*., 2000).

More recently, the K6 was clinically calibrated against the SCID in a larger sub-sample (*N* = 1500), drawn from the US 2008 National Survey on Drug Use and Health (NSDUH, *N* = 45,000), which included major racial-ethnic groups (Aldworth *et al*., 2010; Table 2). The K6 alone predicted SMI (with a lower GAF score of ≤ 50) less accurately than the K6 combined with a reduced 8-item functional impairment scale, the World Health Organization Disability Assessment Schedule (WHODAS; see Table 2). The combined scales provided comparable predictive accuracy across race-ethnic groups (Aldworth *et al*., 2010), although accuracy for the Hispanic group was lower than for the other groups (Table 2). This and a related study (Colpe *et al*., 2010) reported no information on English proficiency inclusion criteria, or translation of the K6.

The final K6 and K10 consisted of six and 10 items, respectively, that asked about symptom frequency in the past month. Differences in scoring methods resulted in differing score ranges: five-point Likert-type response options were scored 0–4 in the US, but 1–5 in Australia, resulting in respective K6 score ranges 0–24 and 6–30, and K10 score ranges 0–40 and 10–50 (Andrews and Slade, 2001; NCS, 2012). In some studies the recall period is changed to the "worst month" in the last 12 months to align with diagnostic instruments.

Although Kessler *et al.* (2002, 2003) recommended the K6/K10 as a useful screening and outcome measure in primary care and other clinical settings, scoring methods and cutoff points for clinical services were left unresolved (Andrews and Slade, 2001). For epidemiological surveys Kessler *et al.* (2002) advised against simple summing of items and cutoff scores as items need to be weighted with values generated from reference (or normative) samples such as national mental health surveys (Kessler *et al.*, 2002). Clinical settings, however, require scoring guidelines, based on clinical norms or reference samples, but these appear to be lacking (Andrews and Slade, 2001). No studies appear to have examined the original K6/K10's predictive validity against diagnostic instruments in primary care or mental health settings. In Australia K10 scoring methods and cutoff scores (or score ranges) for these settings have been based on the epidemiological NSMHWB (Andrews and Slade, 2001; Furukawa *et al.*, 2003).

The only clinically-based predictive validity study identified using the original K6/K10 was with Australian injecting drug users (Hides *et al.*, 2007), with whom a high K10 cutoff score of ≥ 27 adequately predicted DSM-IV affective disorders on the Mini-International Neuropsychiatric Interview (MINI, Sheehan *et al.*, 1998). This finding would be of limited relevance to primary care practices. A study (Haller *et al.*, 2009), reported to have found good K6 concordance in primary care attenders (Kessler *et al.*, 2010), did not validate the K10 against a diagnostic instrument, and showed poor association between K10 scores and general practitioners' (GPs') identification of mental illness. In the apparent absence of standardized clinical guidelines, it remains unclear how primary care providers score and interpret the K6/K10. The paucity of studies on the original K6/K10's predictive validity in primary care contrasts with 11 such studies in NES countries reported later.

Studies using translated or culturally adapted K10s

To enable the cultural equivalence of translated/adapted K6/K10's to be evaluated, studies will be expected to show one or more of the following types of equivalence (see Table 1), each of which will be examined in turn. The sequencing is consistent with steps that would logically be followed if all aspects of cultural equivalence were tested: (a) conceptual and linguistic equivalence; (b) method equivalence; (c) structural equivalence; (d) item and scalar equivalence; and (e) criterion equivalence. Normative equivalence is evaluated throughout. Some studies have examined more than one type of equivalence.

Conceptual and linguistic equivalence

Equivalence in meaning, or conceptual and linguistic equivalence, is fundamental to the cultural equivalence of a translated instrument. To ensure the same construct is measured across cultures, item content should be as familiar and appropriate in the target as in the source language (Harkness *et al.*, 2008). Conceptual and linguistic equivalence is demonstrated by undertaking a comprehensive translation process, such as proposed by Harkness *et al.* (2008) for the World Mental Health Survey (WMHS). In brief, guidelines for the WMHS were: (a) forward- and back-translation by independent expert bilingual clinicians who are native speakers of the target and source languages, respectively; (b) review of each translation by a bilingual expert panel; (c) pre-testing and cognitive interview of representative respondents, who are prompted to "think aloud" to explain their responses; and (d) documentation of cultural adaptations (Harkness *et al.*, 2008).

Studies listed in Table 2 show that the K6/K10 was translated into 26 clearly specified languages, in addition to Spanish (Grzywacz *et al.*, 2009), and Farsi (Sulaiman-Hill and Thompson, 2010). Languages into which the K6 was translated for the 14 countries in the WMHS were not detailed in the international comparative report on the K6 (Kessler *et al.*, 2010), but each country performed its own translation, following WMHS guidelines (Harkness *et al.*, 2008). Three studies have published more detail on WMHS findings in Japan (Furukawa *et al.*, 2008), New Zealand (NZ; Browne *et al.*, (2010), and South Africa (Andersen *et al.*, 2011), discussed further later.

Thirteen reviewed studies cited the official World Health Organization (WHO, 2012) protocol, or variations thereon (Patel *et al.*, 2008; Fernandes *et al.*, 2011), which equates to the WMHS guidelines. An expert consensus approach, described later, was used by four studies. As some languages lack conceptual equivalents for English mental health concepts, reports would be expected of difficulties that might have been encountered in translation, or in reconciling local idioms of distress (Okawa, 2008). Only Grzywacz *et al.* (2009) and Tesfaye *et al.* (2010) provided a comprehensive account of their translation procedures, with other studies providing minimal or abbreviated accounts.

Tesfaye *et al.* (2010), in an urban Ethiopian post-natal depression study, fully documented translation difficulties, illustrating respondents' lack of familiarity with K6/K10 concepts. Steps equivalent to the WHO protocol were followed, with forward and back-translation by independent bilingual psychiatry residents. To examine content validity of the Amharic translation, an interviewer administered the K6/K10 and prompted respondents to explain their answers. Tesfaye *et al.* (2010) reported that: the Amharic translations of “feel depressed” and “worthless” were not understood; in the item “feel tired for no good reason” the term “for no good reason” was ignored and endorsed when respondents had cause to feel tired; and “everything was an effort” was misunderstood as being compelled to act. Issues regarding item connotation are evident for the “tired” and “effort” items. Tesfaye *et al.* (2010) viewed these items as improving detection of post-natal depression because they enabled expression of somatic idioms of distress. However, Fernandes et *al.* (2011), in a study at a rural Indian pre-natal clinic, argued that high levels of endorsement of these items could reflect “normal [pregnancy-related] physiological changes” (p. 210) that may be misdiagnosed as depression, but also could signify anaemia and malnutrition. Although Tesfaye *et al.* (2010) concluded that the Amharic K10 had “good validity” (discussed under criterion equivalence), it was acknowledged that some translated K6/K10 items did not achieve conceptual equivalence, and qualitative studies were needed to explore “the most appropriate idioms of distress in this setting” (p. 107).

Questions regarding connotations of translated items emerged incidentally in some studies. Donker *et al.* (2010a) translated the K10 into Dutch for a primary care study, citing the WHO protocol. Changes in connotation appear to have occurred in translation, as “commonly used Dutch synonyms” for “depression” and for “worthless” (Donker *et al.*, 2010a, p. 47) were added to these items, but not further discussed.

Of five studies that used the expert consensus approach (Table 2), two (Sulaiman-Hill and Thompson, 2010; Laube, 2010) used six of the 14 K10 translations, translated by, and available on the website of the New South Wales (NSW) Transcultural Mental Health Centre (TMHC, 2012). As with the WHO protocol, the K10 was forward- and back-translated; consumer groups were consulted and consensus changes incorporated; and professional health interpreters reviewed final translations (Dr Roy Laube, TMHC, personal communication, 14 October 2011). The TMHC’s Arabic, Chinese, Greek, Italian, and Vietnamese translations were field-tested with NES groups, and used in NSW health surveys (Public Health Division, 2000; Boufous *et al.*, 2005).

The TMHC’s K10 Farsi translation was used in a mental health study of Afghan ($n = 90$) and Kurdish ($n = 103$) refugees (Sulaiman-Hill and Thompson, 2010). The “feel worthless” item was viewed as “culturally problematic”, by some Kurdish respondents, “as it challenged their ideal of human dignity” (Sulaiman-Hill and Thompson, 2010, p. 243). Whether this response was to the Farsi or English K10 was unclear.

The “worthless” item was also considered to have pejorative connotations in a study of working conditions of 40 Latino Spanish-speaking farmworkers (Grzywacz *et al.*, 2009). An expert team approach was used to translate the K6 into Spanish. To investigate the K6’s conceptual and linguistic equivalence cognitive interviews were conducted with the farmworkers. The translated K6 items were viewed as too long and complex, and the language “too formal and intimidating” (Grzywacz *et al.*, 2009, p. 133). The “restless or fidgety” item was perceived as applying to children’s behaviour. Grzywacz *et al.* (2009) questioned the suitability of the K6 for investigating farmworkers’ mental health.

In a French Emergency Department, a French version of the K6/K10 from a Canadian Community Health Survey was used by Arnaud *et al.* (2010), citing a French-language reference for the translation. A table listing the K6/K10 items (and factor loadings) raises questions about the translation, as “hopeless” appears to have been back-translated into English as “despairing”, and “worthless” as “good for nothing” (Arnaud *et al.*, 2010, p. 1239).

No cultural adaptations were required to the K6 for American Indian communities as consulted members expressed no concerns about its cultural validity (Mitchell and Beals, 2011). However, cultural adaptations were made to the K6 following consultation with Indigenous Australians prior to a mental health survey. The “worthless” item was omitted as it “might be considered offensive” (AIHW, 2009, p. 5) to Indigenous respondents, resulting in the K5. To improve Indigenous Australians’ understanding, the “feel hopeless” item was changed to “feel without hope”, and “restless or fidgety” to “restless or jumpy” (AIHW, 2009). In the NZ WMHS Maori and Pacific people were consulted and participated in the project (Wells *et al.*, 2006; Browne *et al.*, 2010), but no reports were provided on possible K10 cultural adaptations.

Subtle changes in item connotation that may occur in translation (e.g. Arnaud *et al.*, 2010) or in different cultural contexts (e.g. AIHW, 2009) may result in differential item functioning (DIF). DIF may occur when cultural groups have a different probability of endorsing an item, obtaining differential mean scores on individual items, and potentially impairing scalar equivalence (Huysamen, 2002). DIF is discussed further under item and scalar equivalence.

Conceptual and linguistic equivalence is difficult to evaluate in the majority of reviewed studies as direct reports on translation issues are lacking. Nevertheless, misunderstanding of some items, and changes to ensure understanding or to avoid offense, raise questions regarding item relevance and appropriateness, and hence conceptual equivalence.

Method equivalence

Method equivalence requires that test manuals be provided for translated/adapted instruments to ensure cultural equivalence in administration procedures, with advice on how to explain instruments and their purpose, and how to respond to questions (Nell, 2000). To contribute to a test's cultural equivalence, administrative procedures have the same meaning, and are not differentially influenced by sample differences, cultural norms, stimulus familiarity, poor translation, or communication failure (Van de Vijver and Poortinga, 2005). Guidelines are included on engaging interpreters or bilingual professionals if the respondent is not proficient in the host country language, even if the test is translated (AERA *et al.*, 1999). Test manuals provide scoring procedures that enable professionals to quantify and interpret scores (AERA *et al.*, 1999; Van de Vijver and Leung, 2011).

For primary care and mental health providers, guidance on all these administration aspects is important to ensure that culturally diverse patients understand the test and respond as intended. Patients may not speak English, may be illiterate, unfamiliar with mental health instruments and, if they originate from countries lacking a "tradition of free speech" (WHO WMHS Consortium, 2004, p. 2587), may be reluctant to complete official forms and admit to emotional problems (AERA *et al.*, 1999; Huysamen, 2002).

No standardized administration guidelines have been located for the K6 or K10 for English or non-English versions. Studies vary in whether the K6/K10 is self-completed, or administered by an interviewer, in person or by telephone. Deference to the interviewer may influence responses (Van de Vijver and Leung, 2011), but no studies seem to have investigated influences of K6/K10 completion style, which would be important in clinical settings. Method inequivalence may influence scores and contribute to significant differences in mean scores, or to individual elevated or low scores that are not due to the construct under investigation (Van de Vijver, 1998). Refugees may give acquiescent or socially desirable responses, showing extreme response styles, under- or over-reporting levels of distress to ensure refugee status or receipt of services (Johnson, 1998; Loutan *et al.*, 1999). Test bias is likely to increase with the "cultural distance to be bridged" by an instrument (Van de Vijver, 1998, p. 43).

As K6/K10 translation/adaptation studies vary in item coding methods used (0–4, or 1–5) confusion can be caused when comparing scores across countries (Fernandes *et al.*, 2011). Patel *et al.* (2008) made K6/K10 response options dichotomous to facilitate completion. Pre-natal women with low literacy in rural India were unfamiliar with K10 Likert-type response scales, necessitating administration by an interviewer (Fernandes *et al.*, 2011).

There are also widely varying methods for calculating total K6/K10 scores that appear to compromise scalar equivalence, thereby precluding cross-cultural score comparisons. Space forbids full discussion of scoring methods, but they have included optimal maximum-likelihood estimates of true psychological distress; summing sensitivities for endorsed items (Kessler *et al.*, 2002); stratum specific likelihood ratios (SSLRs; Furukawa *et al.*, 2003; Slade *et al.*, 2011); and multiple imputation estimation methods to predict the probability of SMI (Kessler *et al.*, 2010). These methods tend to be "computationally demanding" (Sunderland *et al.*, 2011, p. 888) and impractical in clinical settings, where immediate clinical decisions are required. Although simple summing of items was not recommended (Kessler *et al.*, 2002), translation studies appear to do so (e.g. Baggaley *et al.*, 2007; Andersen *et al.*, 2011) or make no reference to item weights or score calculation (e.g. Patel *et al.*, 2008).

Studies also vary in whether the original 30-day recall period is used, or the worst month in the last 12 months, to align with a diagnostic instrument. Browne *et al.* (2010) in the NZ WMHS found that the K10 worst month recall period better predicted diagnosis than the past month. However, Kessler *et al.* (2010) found no difference between the two recall periods for prediction from the K6 in the WMHS, with high correlations between the two periods.

Wide variations in the way that the K6/K10 is administered and scored suggest that cutoffs and means cannot justifiably be compared across cultures. For clinicians working with culturally diverse clients, these method variations also limit conclusions that can be drawn from cross-cultural K6/K10 studies.

Structural equivalence

Evidence for the structural equivalence of translated/adapted K10s is reviewed before criterion equivalence, as judgements about caseness are compromised if the underlying construct differs across groups (Van de Vijver and Poortinga, 2005). Psychometric methods used to determine whether translated/adapted instruments show identity of underlying dimensions include calculation of internal reliability coefficients and factor analysis (Van de Vijver and Poortinga, 2005). Fourteen studies reported Cronbach alphas, indicating moderate to high reliability (Table 2). In studies that used both scales, alphas tended to be lower for the K6 than the K10. Although high alphas tend to be interpreted as demonstrating an homogeneous underlying construct, high alphas may be obtained when items are heterogeneous (Green *et al.*, 1977). Factor analysis therefore is required to investigate a scale's internal structure, and to establish whether the underlying construct is stable across culturally diverse groups (Van de Vijver and Poortinga, 2005).

Only one clinical and six epidemiological studies have been identified that factor analysed translated/adapted K6/K10 data, results of which are inconsistent. Four studies supporting Kessler *et al.*'s (2002) original K10 unidimensional model are reviewed first. Structural equivalence with the original K6 was demonstrated by exploratory factor analysis of K6 data from the 14 WMHS countries, which found support for a single factor (Kessler *et al.*, 2010). In a Japanese survey of complicated grief, exploratory factor analysis of combined items from the K6 and the Brief Grief Questionnaire found no cross-loading of items from the two scales, with K6 items loading onto a single factor (Ito *et al.*, 2012). Confirmatory factor analysis of K6 data from American Indian communities showed that a single factor provided a satisfactory fit (Mitchell and Beals, 2011).

In the Netherlands Fassaert *et al.* (2009a) conducted exploratory and confirmatory factor analysis of K10 combined group data from Dutch, Moroccan and Turkish groups. A single factor provided a sufficient fit across the three groups (Fassaert *et al.*, 2009a). However, as discussed in the next section, Fassaert *et al.* (2009a) found significant ethnic DIF for three items.

The unidimensional model was not supported by three translation studies, one of which was clinically based. In a French Emergency Department Arnaud *et al.* (2010) randomly allocated the K6 and K10 to patients with alcohol-related disorders. Exploratory factor analysis identified two K6 and three K10 factors; a single item ("nervous") loaded on the third factor. The factors were not interpreted by Arnaud *et al.* (2010) but K6 item loadings and items loading on the first two K10 factors appeared consistent with depression and anxiety. This study seems limited by small numbers completing each scale (Table 2). In a Hong Kong survey exploratory factor analysis of K6 data showed that a two-factor solution (depression and anxiety) provided the best fit (Lee *et al.*, 2012).

Using five of the TMHC's K10 translations, K10 health survey data were obtained from six language groups in Australia (Laube, 2010; Table 2). Exploratory and confirmatory factor analysis showed that two factors provided a better fit in each group, suggesting structural inequivalence with the original K10. Item-to-factor loadings differed between groups and were difficult to interpret: no pattern could be identified that might be shared by, e.g. Greek and Italian data, or Chinese and Vietnamese data. However, a pattern that could be categorised as 'emotional' and 'somatic/behavioural' emerged from the Arabic and Italian data (Laube, 2010). As the K10's latent constructs were not identical across language groups, Laube (2010) concluded that expressions of psychological distress varied across cultural communities, and K10 items may not represent important aspects of NES respondents' psychological distress.

Research with the original K10 has also questioned the unidimensional factor structure. Laube's (2010) emotional and somatic behavioural factors resembled depression and anxiety factors found in K10 data from an Australian survey (Brooks *et al.*, 2006), and in US and Australian clinical data (O'Connor *et al.*, 2012; Sunderland *et al.*, 2012a). However, a unidimensional model provided a good fit for Australian K6 clinical data and for K6 and K10 Australian survey data (Sunderland *et al.*, 2012a). It would appear that the internal structure of the K6/K10 remains an open question in both clinical and survey settings.

Item and scalar equivalence

DIF occurs when different cultural groups that have the same position on an underlying latent trait (e.g. psychological distress), have a different probability of endorsing an item (Huysamen, 2002; Van de Vijver and Poortinga, 2005). Findings of DIF indicate that group membership may be influencing item endorsement rather than the construct's latent dimension (Sunderland *et al.*, 2012b). Although total scale scores may be similar, DIF analysis may show that cultural groups obtained differing mean scores on individual items, resulting in item and scalar inequivalence (Huysamen, 2002; Van de Vijver and Poortinga, 2005). No translation/adaptation studies have been identified that reported item-level data on DIF, comparable to Sunderland *et al.* (2012b), who showed significant age-related DIF on K6 items.

Only Fassaert *et al.* (2009a) directly investigated DIF. After establishing the unidimensionality of the Dutch, Moroccan and Turkish K10s, Fassaert *et al.* (2009a) reported that Dutch respondents were significantly more likely to endorse "restless or fidgety", while Turkish and Moroccan respondents were more likely to endorse "everything was an effort". "Feel tired out for no good reason" also showed significant DIF, but this was judged "not relevant" (Fassaert *et al.*, 2009a, p. 164), without further explanation. These findings raise doubts whether the two translated K10s achieved conceptual and linguistic equivalence: the K10 was translated into Turkish using only forward and back-translation, and Moroccan interviewers translated directly from a pre-existing Dutch K10, as Moroccan-Arabic and Berber are not written languages (Fassaert *et al.*, 2009a).

Scalar equivalence assumes that tests have the same measurement units and that total scores can be compared across cultures (Van de Vijver and Poortinga, 2005). Method and item inequivalence, which we have suggested may occur for translated/adapted K6/K10s, can impair scalar equivalence. Mean differences between cultural groups therefore should not be taken at face value, but factors influencing scalar equivalence, such as response styles and changes in item connotation should be ruled out (Huysamen, 2002). A number of K6/K10 studies have reported significantly higher or lower mean scores than the majority population for Indigenous (AIHW, 2009; Browne *et al.*, 2010), immigrant (Boufous *et al.*, 2005; Fassaert *et al.*, 2009b), refugee (Sulaiman-Hill and Thompson, 2010), racial, ethnic, and language (Albrecht and McVeigh, 2012) groups. The meaning of these differential scores is unclear due to potential scalar inequivalence. Moreover, only one of these studies (Fassaert *et al.*, 2009a, 2009b) clinically validated the K6/K10 for the groups under investigation. For the other studies, relevant normative databases or cutoff scores against which scores could be interpreted were lacking (Sulaiman-Hill and Thompson, 2010).

Criterion equivalence: case prediction

To demonstrate criterion equivalence, translated/adapted K6/K10s should predict cases on diagnostic instruments as accurately as the original K6/K10, shown by comparable predictive validity findings. As noted earlier, no studies were located on the original K6/K10's predictive validity in primary care or mental health settings; consequently comparisons will be drawn with the original predictive validity findings in the clinical reappraisal survey (Kessler *et al.*, 2002, 2003). Predictive validity for SCID-based ADDs was shown by analysis of the receiver operating characteristic (ROC) curve.

ROC analysis provides information on the area under the curve (AUC), and on the relationship between sensitivity (true cases), specificity (true non-cases), and positive predictive values (PPVs; Fassaert *et al.*, 2009a). The AUC represents the probability that "randomly chosen cases and non-cases would be correctly distinguished" based on K6 or K10 scores (Kessler *et al.*, 2002, p. 966). An AUC of 1.0 denotes perfect predictive accuracy; ≥ 0.80 represents good accuracy; while 0.5 represents chance detection. PPVs indicate the proportion of true positives, but unlike sensitivity and specificity, are dependent on population prevalence of the disorder. A PPV ≥ 50 reduces the number of false positives, but risks reducing sensitivity. Optimal K6/K10 cutoff scores for predicting cases on a diagnostic instrument are chosen by balancing trade-offs between sensitivity, specificity and PPVs (Lalkhen and McCluskey, 2008; Patel *et al.*, 2008; Fassaert *et al.*, 2009a), hereafter abbreviated as ROC measures or values. Some studies also base cutoffs on the percentage of cases correctly classified: the proportion of the selected population correctly classified, as true positives and true negatives (Zhu *et al.*, 2010).

The K6 and K10 showed high predictive accuracy for SCID-based ADDs with respective AUCs of 0.879 and 0.876 (Kessler *et al.*, 2002; Table 2). For prediction of SMI (defined as a DSM-IV disorder with a GAF score ≤ 60) the K6 showed low sensitivity of 0.36 and high specificity of 0.96 at a cutoff score of ≥ 13/24 (Kessler *et al.*, 2003). With a lower GAF score of ≤ 50, Aldworth *et al.* (2010) showed similar sensitivity and specificity at a ≥ 17/24 K6 cutoff (Table 2). For the K10 Kessler *et al.* (2002) plotted sensitivity against 1-specificity but did not report sensitivity and specificity for specific K10 scores. However, in the NSMHWB, Andrews and Slade (2001) found balanced sensitivity of 0.81 and specificity of 0.83, at a K10 cutoff of ≥ 17/50 (Table 2).

The original K6's sensitivity in the vicinity of 0.36 (Kessler *et al.*, 2003; Aldworth *et al.*, 2010) seems particularly low for the purposes of clinical practice, as this would provide an unacceptable 64% of false positives. Balanced sensitivity and specificity at values ≥ 0.80, such as shown for the K10 in the 2007 NSMHWB (Andrews and Slade, 2001), are to be preferred for clinical practice to minimize clinicians' time with false positive cases, while also reducing the risk of failure to detect cases through false negatives (Donker *et al.*, 2010a).

Eleven clinical and seven epidemiological studies examined accuracy of case prediction and showed widely varying predictive validity for the K6 and K10. The clinical studies may be loosely classed as involving patients at peri-natal (4), primary care (3), mental health related (3), and HIV (1) services, and are reviewed in this sequence.

In two studies of peri-natal depression and common mental disorders (CMDs) reasonably balanced and high sensitivity and specificity were obtained by Fernandes *et al.* (2011) in India and Tesfaye *et al.* (2010) in Ethiopia. Low cutoffs were recommended by both studies to ensure identification of women at risk of depression or physical disorders. A K10 cutoff score of ≥ 6/40 was recommended by Fernandes *et al.* (2011) while Tesfaye *et al.* (2010), recommended a cutoff of six to seven for the K10, and four to five for the K6 (score ranges not reported). Tesfaye *et al.* (2010) noted that there was little difference in the K6 and K10's predictive accuracy.

The low cutoffs recommended by these two studies are at odds with Baggaley *et al.*'s (2007) K6 and K10 cutoffs of ≥ 10/24 and ≥ 14/40, respectively, to detect post-natal depression in Burkina Faso. On the one hand, Baggaley *et al.* (2007) recommended the higher K10 cutoff because it improved the proportion accurately classified, and would minimize allocation of scarce resources to non-cases. On the other hand, Baggaley *et al.* (2007) suggested a lower cutoff in settings where women were at risk of depression. Spies *et al.* (2009) in South Africa, selected an even higher K10 cutoff of ≥ 21.5 (score range not reported) for predicting pre-natal SCID-DSM-IV major depression as this cutoff provided "acceptable" accuracy (Spies *et al.*, 2009, p. 71), However, low ROC values suggest a relatively high misclassification rate of both true cases and non-cases. The varying cutoffs, and the absence in some studies of K6/K10 score ranges highlights difficulties in comparing scores across cultures.

Of three primary care studies (reported next), Donker *et al.* (2010a) in the Netherlands obtained satisfactory and balanced ROC values for predicting CIDI-based ADD at a K10 cutoff of ≥ 20/50, suggesting criterion equivalence with the original K10. Difficulties in choosing cutoffs that balanced ROC values, were shown by the other two studies. Carra *et al.* (2011) in Italy, found that a K10 cutoff 13/40 provided balanced but low ROC values for prediction of SCID-DSM-IV ADDs, so recommended a lower K10 cutoff of 12/40 to ensure detection of true cases, but at the expense of low specificity. Cutoffs on the K10 of 13/40 and on the K6 of 7/24 provided balanced and relatively high sensitivity and specificity for the prediction of SMI (Carra *et al.*, 2011). In contrast to Carra *et al.* (2011), Patel *et al.* (2008) in India recommended higher K6 and K10 cutoffs (3–4/6 and 6–7/10, respectively) than required for optimal accuracy, to reduce demands on busy primary care practices by non-cases. High specificity maximized true negatives, but ensuring a high PPV resulted in low sensitivity, so that one third of cases would be misclassified as false positives. Patel *et al.*'s (2008) predictive accuracy may have been reduced by making K6/K10 response options dichotomous and changing the recall period to two weeks.

Three mental health-related clinical studies (reported next) found that optimal cutoff points yielded notably lower specificity than sensitivity, which would result in between 25% to 33% of false negatives Arnaud *et al.*'s (2010) study of French Emergency Department patients found that optimal K6 and K10 cutoffs of ≥ 10 and ≥ 14 (respectively; score ranges not reported) for detecting a MINI-based diagnosis of alcohol-related disorders yielded high sensitivity but low specificity for both scales. Nevertheless, Arnaud *et al.* (2010) recommended adoption of the K6 because of its brevity. In a small sample of Japanese psychiatric outpatients ($n = 17$), Sakurai *et al.* (2011) found that a 4–5/24 K6 cut-score, and 9–10/40 K10 cutoff provided perfect sensitivity but low specificity for psychiatrist-diagnosed DSM-IV ADD. This study is flawed by inclusion in the ROC calculations of a randomly-selected community sample ($n = 147$), who were assumed to be non-cases and not assessed for diagnosis.

In a Dutch study people were recruited on the Internet, inviting those who were depressed, anxious or had alcohol use problems to complete the K10 (Donker *et al.*, 2010b). A sub-sample of respondents was administered the CIDI by telephone. Although resulting in low specificity, Donker *et al.* (2010b) selected a K10 cutoff of ≥ 29/50 that provided a maximum sum of sensitivity and specificity for predicting CIDI-DSM-IV depressive disorders. This web-based cutoff was notably higher than the K10 ≥ 20/50 cutoff found by Donker *et al.* (2010a) with pen and paper K10s in primary care. The discrepancy was attributed to greater self-disclosure to the internet-based K10 than in the phone-based CIDI interviews, where respondents may have felt less anonymous (Donker *et al.*, 2010b).

A study of people infected with HIV in India, found that a K6 cutoff of ≥ 13/30 provided balanced but relatively low ROC values for the detection of psychiatrist diagnosed ICD-10 (International Classification of Diseases, 10th revision) CMDs (Chowdhary and Patel, 2010). With 0.54 specificity, almost half of true cases would not be detected.

Seven epidemiological studies examined predictive validity. This includes a comparative report on the K6 in 14 countries participating in the WMHS (Kessler *et al.*, 2010). In the WMHS, following corrections for differential sensitivity associated with age, gender and education in some countries, differing formulae predicting CIDI-DSM-IV SMI were estimated for each country, using various logistic regression equations, and the multiple estimation method (Kessler *et al.*, 2010). Sunderland *et al.* (2011) used this method with Australian 2007 NSMHWB K6/K10 data and showed that a quadratic form of the K6, controlling for age, (and for the K10 also controlling for gender) best predicted CIDI-DSM-IV SMI, but ADDs were predicted less accurately. In the WMHS, AUCs for prediction of SMI from the K6 ranged from 0.76 for South Africa, to 0.86 for Lebanon, with a median AUC of 0.83 (Kessler *et al.*, 2010). Sensitivity and specificity values were not reported. Kessler *et al.* (2010) noted that the K6's primary value is as a "broad screener" (p. 17) for SMI, rather than for particular disorders. In addition, the median AUC of 0.83 would still leave 17% of cases undetected (Kessler *et al.*, 2010).

More detailed K6/K10 findings were reported by three WMHS studies (reported next). Furukawa *et al.* (2008) considered the K6/K10's performance in a Japanese health survey to be "essentially equivalent" (p. 157) to the original K6/K10. Rather than K6/K10 cutoff scores, Furukawa *et al.* (2008) calculated SSLRs. The likelihood of a disorder may be ruled out, or in, with SSLRs of < 0.1 or > 10, respectively (Furukawa *et al.*, 2003). An SSLR of 16 on the Japanese K6 showed there was an increased odds of a CIDI-DSM-IV ADD in the K6's 9–13/24 score range. On the K10 respective scores were an SSLR of 11 for an increased odds of a disorder in the K10's 15–19/40 score range. These SSLRs were comparable to those found in the Australian NSMHWB (Furukawa *et al.*, 2008).

Furukawa *et al.*'s (2008) findings contrast with the South African WMHS (Andersen *et al.*, 2011). Andersen *et al.* (2011) reported that in combined racial/ethnic group (Black and Other; see Table 2) data no K6 or K10 cutoff could be identified, that provided both acceptably balanced sensitivity and a PPV ≥ 50, for predicting CIDI-DSM-IV ADD. To obtain, a PPV ≥ 50, K10 sensitivity would be 4% at a cutoff of 42/50. In addition, the K6 and K10 predicted ADD significantly less accurately in the Black than in the Other group (Table 2), which Andersen *et al.* (2011) attributed to possible cultural differences in symptom expression. Severe economic disadvantage also may have caused Black respondents to endorse the "effort" and "worthless" items, regardless of the presence of mental illness, possibly suggesting DIF (Andersen *et al.*, 2011). SMI was not examined by Andersen *et al.* (2011) but Kessler *et al.* (2010) found that, of the 14 WMHS countries, the K6 in South Africa showed the lowest AUC (0.76) for the prediction of SMI. Andersen *et al.* (2011) concluded that the K10 was unsuitable for South African clinical or epidemiological settings until further research was conducted into DIF and clinical calibration of the K10.

For the NZ WMHS Browne *et al.* (2010) did not report on the K6, but the K10 predicted CIDI-based DSM-IV ADD marginally less accurately than SMI. Maori and Pacific people showed significantly higher adjusted K10 mean scores than the combined Other ethnic group, but K10 cutoffs, sensitivity and specificity for the whole sample were not reported, and no ROC data were provided for ethnic groups.

Fassaert *et al.*'s (2009a) Dutch health survey, found somewhat low but comparably balanced sensitivity and specificity for Dutch, Moroccan and Turkish groups. However, this required higher cutoff scores for the immigrant groups, and was at the cost of low PPVs for the Dutch and Moroccan groups, suggesting the likelihood of a high proportion of false positives. Fassaert *et al.* (2009a) did not report K6 results but noted that intercorrelations between K10 items indicated redundancy, and recommended the K6 for future use.

In a Hong Kong mental health survey, a K6 cutoff of 12–13/24 to predict SCID-DSM-IV SMI provided a low AUC of 0.69 (Lee *et al.*, 2012). Although Lee *et al.* (2012) concluded that the Chinese K6 was a valuable screening measure for SMI in epidemiological surveys, high specificity and very low sensitivity and PPV made the K6 "a better screen-out than screen-in tool of SMI" (p. 590). A health survey of American Indian communities showed that the frequently-used K6 cutoff of ≥ 13 resulted in very low sensitivity, but high specificity for any CIDI-DSM-IV mood disorder (Mitchell and Beals, 2011). Therefore a lower cutoff might be preferred to increase the probability of "true caseness" (Mitchell and Beals, 2011, p. 759).

No diagnostic instrument was used to validate the K5 in Australia's National Aboriginal and Torres Strait Islander Health Survey (NATSIHS) but K5 scores showed convergent validity with self-reported mental illness, and with other stressors including health, racial discrimination, unemployment and separation from family (AIHW, 2009). Of Indigenous respondents with low K5 scores, 7% self-reported mental illness, compared with 33% of respondents with very high K5 scores (AIHW, 2009; Cunningham and Paradies, 2012; Table 2).

In the only study identified that used a Kessler scale and a diagnostic instrument with a refugee-like population, the K10's predictive validity could not be evaluated. In a mental health survey of the internally displaced people of Timor Leste, Silove *et al.* (2008) used a K10 cutoff of ≥ 30 (score range not reported), presumably to identify severe mental illness. No K10/SCID predictive values were reported, but 82% of SCID-based non-psychotic disorders were screen positive on the Harvard Trauma Questionnaire and/or K10 (Silove *et al.*, 2008). Separate results for the two screening measures were not reported.

No clear pattern emerges from the predictive validity studies of translated/adapted K6/K10s, except that ROC values in studies that reported findings for both the K6 and K10 did not show marked differences. Otherwise, widely varying scoring methods, cutoff scores, and ROC values in different countries, do not provide strong evidence for criterion equivalence comparable to the original K6 and K10. Relatively high rates of misclassification of true cases and non-cases raise questions as to how well the K6/K10 applies to non-Western groups. Interpretation of results is confounded by differences in sample types, sample size, and criterion diagnostic measures.

Sensitivity to change

Although adopted as an outcome measure in primary care and MHSs (Hickie *et al.*, 2002; Sunderland *et al.*, 2012a) no publications have been located on translated K6/K10s' sensitivity to change. However, the original K10 has been used as an outcome measure with Indigenous patients (Nagel *et al.*, 2009; Mathieson *et al.*, 2012). Primary care providers delivered culturally modified brief cognitive behavioural therapy (CBT) to 16 Maori patients with K10 scores < 30 (Mathieson *et al.*, 2012). K10 scores showed non-significant improvements at follow-up. The study was limited by small numbers and the absence of a control group (Mathieson *et al.*, 2012). Nagel *et al.* (2009) evaluated a culturally adapted motivational programme for 49 Australian Indigenous patients diagnosed with mental illness and substance dependence. The abstract stated that scores on the Health of the Nation Outcome Scales (HoNOS) and K10 improved significantly following the intervention. However, the article showed HoNOS, but not K10 score changes.

Space limitations prevent full review of sensitivity to change studies with ES participants, which generally excluded those with low English proficiency. However, evidence for the K6/K10's sensitivity to change is limited or methodologically flawed. Minimal K10 score changes were shown following CBT interventions in primary care (Hickie *et al.*, 2010) and at an anxiety clinic (Perini *et al.*, 2006). A study of employee well-being showed only indirect effects on K6 scores of changes in flexibility of work hours (Moen *et al.*, 2011). Hides *et al.* (2011) failed to interpret K10 change data following CBT intervention for young people with comorbid depression and substance misuse. In an evaluation of the Better Outcomes in Mental Health Care Programme, K10 scores improved significantly, but findings were flawed by lack of independent K10 data collection by treating mental health professionals and GPs (Pirkis *et al.*, 2010; Allen and Jackson, 2011). Significant K10 score changes were found by Stallman *et al.* (2010): women who had miscarried showed significantly lower K10 scores at three-months' follow-up than at miscarriage.

The K10+ (four additional unscored items assess functioning and disability) was adopted as a consumer-rated measure by MHSs in NSW and South Australia when collection of outcome measures became mandatory under Australian national mental health policy in 2003 (Trauer, 2010). Since then MHS have published little data on outcome measures, with the exception of the clinician-rated HoNOS (Trauer, 2011). No K10+ outcome research has been located with either ES or non-ES mental health patients. A more comprehensive review would consider the sensitivity of the K6/K10 against other change measures to determine whether the K6/K10 lacked sensitivity or the interventions were ineffective.

Discussion

Development of the original K6/K10 followed a rigorous psychometric methodology ensuring its construct validity in English-language surveys. The K6's low sensitivity and high specificity in these settings however, appear to make it more suitable as a rule-out rather than a rule-in instrument for SMI (Lee *et al.*, 2012). The K6/K10's cultural equivalence for race-ethnicity was tested in the US for Hispanic and Black groups (Kessler *et al.*, 2002), for whom no K6/K10 validity data were published, and more recently for Hispanic, Black and Other ethnic groups (Aldworth *et al.*, 2010) for whom validity data were published for the K6 combined with the eight-item WHODAS, but not for the K6 alone. English proficiency inclusion criteria were not reported, potentially excluding, for example, the 29% of the Spanish-speaking population who spoke English not well, or not at all (Shin and Kominski, 2010). A finding of significantly higher K6 scores for NES than ES speakers in a New York Community Health Survey, led Albrecht and McVeigh (2012) to recommend further research into "non-English language versions of the K6" (p. 5). This study did not report translation procedures or validation of the K6 for its culturally diverse population.

The original K6/K10's applicability in primary care and mental health services is uncertain as no predictive validity studies with the original K6/K10 were identified in these settings, resulting in a lack of clinical norms (Andrews and Slade, 2001), and raising concerns about the accuracy of case identification by clinical practitioners. The K10 appears to be more widely used in Australia, both clinically (Pirkis *et al.*, 2010), and epidemiologically (Slade *et al.*, 2011), while both the K10 and K6 were used in non-ES countries (Table 2). In the US the K6 appears to be preferred (e.g. Aldworth *et al.*, 2010), but does not appear to have wide clinical usage.

Some publications state that the K6/K10 has been validated in various languages (Carra *et al.*, 2011; Slade *et al.*, 2011), but data in the citations do not necessarily support these claims, as studies using K6/K10 translations/adaptations provide inconsistent evidence of cultural equivalence. One such cited study (Tesfaye *et al.*, 2010) highlighted difficulties in achieving conceptual and linguistic equivalence in Amharic. Across nine studies, DIF, or suggestions of DIF were reported for a total of six K10 items, including: "tired for no good reason", "hopeless", "restless or fidgety", "depressed", "everything was an effort", and "worthless".

Some groups' reactions of offense at the "worthless" item (AIHW, 2009; Grzywacz *et al.*, 2009) illustrate the importance of investigating changes in cultural connotations resulting from translation. Black South Africans' endorsement of the "effort" item was attributed to socioeconomic disadvantage (Andersen *et al.*, 2011), which may also explain DIF on this item by immigrant groups in the Netherlands (Fassaert *et al.*, 2009a). However, qualitative analysis of high scores on a similar item by migrants in Canada showed that "effort" was perceived as necessary endeavour, rewarded by achievements in the host country (Moreau *et al.*, 2009). Differing explanations by Fernandes *et al.* (2011) and Tesfaye *et al.* (2010) of peri-natal women' responses to the "effort" item further demonstrate the need to investigate respondents' understanding of items, to ensure conceptual equivalence. Novak *et al.* (2010) argue that "DIF does not imply a poorly measured latent trait" (p. 58), but that sub-group analysis is needed to establish whether culturally diverse groups might need different cutoff scores.

As no standardized manual or guidelines have been located for administering the K6/K10, and for scoring and interpreting scores with either majority or culturally diverse respondents, administrative and method equivalence cannot be ensured. Method and scalar equivalence are compromised by differing methods for scoring items (0–4 and 1–5), and for calculating total scores. These differences prevent comparison of findings across cultures, and provide little in the way of guidance to clinical practitioners.

Evidence for the structural equivalence of translated K6/K10s is inconsistent. Kessler *et al.*'s (2002) original unidimensional structure was supported by four epidemiological studies (including the WMHS), but no clinical studies. In the context of questions regarding the "effort" item's conceptual equivalence it is of interest that this item formed a unique second factor in the Indian WMHS; however, its loading of 0.50 on the first factor was considered acceptable (Kessler *et al.*, 2010). A two-factor model was found by one clinical (Arnaud *et al.*, 2010), and one epidemiological study (Lee *et al.*, 2012). Another epidemiological study found two factors within each of six language groups, but factor structures differed across groups (Laube, 2010). If the K10's psychological distress construct differs across cultural groups, judgements about caseness may be compromised (Van de Vijver and Poortinga, 2005). A two-factor model, however, would be consistent with Kessler *et al.*'s (2002) original division of K6/K10 items into those related to depression and anxiety diagnoses, and potentially provides clinicians with a "richer clinical picture than a single severity score" (Brooks *et al.*, 2006, p. 68), although the sub-scales should not be considered as diagnostic.

Evidence for criterion equivalence of translated/adapted K6/K10s also is equivocal. Cutoff scores for prediction of ADDs varied from 4–5/24 to 13/24 for the K6 and from 6/40 to 29/50 for the K10 (Table 2), but differing minimum scores and failure by some studies to report K6/K10 score ranges, made cutoff scores difficult to interpret and compare. Of 11 clinical studies, three achieved balanced and moderate to high sensitivity and specificity in excess of 0.80, thereby reducing the risk of false positives and false negatives. Furukawa *et al.*'s (2008) epidemiological study found SSLRs comparable to the original K6/K10. Nine studies compromised on accuracy when selecting K6/K10 cutoff scores: five clinical studies chose cutoffs that provided high sensitivity, at the expense of specificity; conversely, in two clinical and two epidemiological studies the cutoffs resulted in higher specificity than sensitivity. With both approaches there would be an increased likelihood of misjudgements about caseness or non-caseness. Aldworth *et al.* (2010) point out that where the K6/K10 yields a high number of false-positives and/or false-negatives the scales have "limitations when compared with a direct … clinical interview" (p. 79).

Andersen *et al.*'s (2011) epidemiological study could find no satisfactory cutoff and considered the K6/K10s translated for the South African WMHS unsuitable for epidemiological or clinical purposes. Fassaert *et al.*'s (2009a) conclusion that the K10's construct was "invariant across three ethnic groups" (p. 166) did not seem justified, given DIF on three items, low PPVs for two groups, and the need for higher cutoffs for the immigrant groups.

In eight studies, use of the CIDI (Table 2) as a diagnostic criterion may have compromised the K6/K10 predictive validity findings, as the CIDI has not been validated in non-Western settings (WHO WMHS Consortium, 2004; Fassaert *et al.*, 2009a). This does not apply to Mitchell and Beals' (2011) study, as the CIDI was culturally adapted for American Indian communities. Despite its low 29% sensitivity, the K6 was considered acceptable to American Indian communities and recommended for research and clinical purposes

(Mitchell and Beals, 2011). Australian Indigenous stakeholders found the K5 acceptable in the NATSIHS and recommended including the full K10 in future surveys (AIHW, 2009).

When considered as two groups, neither clinical studies nor epidemiological studies consistently showed better structural or criterion equivalence, which suggests that translated/adapted K6/K10s were not better suited to clinical or epidemiological settings. Doubts about cultural equivalence are highlighted by findings of studies that included multicultural samples. Of eight studies that included multiple linguistic or ethnic groups (three clinical, five epidemiological) only three conducted group comparisons, but each found group differences in K6/K10 structural (Laube, 2010), or criterion validity (Fassaert *et al.*, 2009a; Andersen *et al.*, 2011).

A number of studies found significant mean differences in K6/K10 scores between culturally diverse groups (e.g. Albrecht and McVeigh, 2012). High scores may well be indicative of high levels of psychological distress, given Indigenous people's experiences of dispossession and economic disadvantage (Vos *et al.*, 2009; Gone and Trimble, 2012), and refugees' and migrants' pre-arrival and post-settlement stressors (Davidson *et al.*, 2008; Kirmayer *et al.*, 2011). However, mean differences should be investigated to rule out cultural factors potentially contributing to DIF (Huysamen, 2002). Given the limitations of the instruments, and the absence of norms, or clinically validated cutoff scores, mean differences can be difficult to interpret.

No studies were identified that examined sensitivity to change with K6/K10 translations. Despite the K6/K10's widespread use in primary care and mental health settings, the suitability of the K6/K10 as an outcome measure with culturally diverse groups is essentially untested. Limited evidence has been shown for the original K6/K10's sensitivity to change.

For clinicians in multicultural societies the cultural equivalence of translated K6/K10s cannot be taken for granted, and the research provides little in the way of guidance regarding appropriate cutoff scores for culturally diverse patients. Discrepancies in recommended cutoffs indicate the importance of first validating and establishing norms in target populations (Fernandes *et al.*, 2011). When cutoff scores have not been validated in a local population it is difficult to interpret the significance of an individual's scores in a clinical setting. Suggestions that higher cutoffs be used for culturally diverse patient groups, require clinical norms for majority populations as a reference point, and even these appear to be lacking.

If the K6/K10 is used to screen culturally diverse patients, high-level scores should be taken seriously as they may be indicative of mental health problems. High or low scores should not be taken at face value however, but should be followed up by a culturally sensitive clinical interview to ascertain the significance of scores for assessment and diagnostic purposes (Huysamen, 2002).

The following research recommendations arise from this review: establish the K6/K10's conceptual equivalence using qualitative research methods (similar to those used by Tesfaye *et al.*, 2010), to examine culturally diverse patients' understanding of translated/adapted K6/K10 items, and clinicians' interpretation of scores; investigate structural and criterion equivalence with clinical data from both the original and translated/adapted K6/K10s; establish clinical norms for majority and culturally diverse populations to ensure accurate norm-referenced K6/K10 score interpretations; develop clinical manuals or guidelines for culturally appropriate administration and scoring; and conduct studies of sensitivity to change. As cutoffs and norms may be difficult to establish for all cultural groups in multicultural societies (Huysamen, 2002), the K6/K10 should be used as intended: as a screening tool, not a diagnostic instrument, with follow-up clinical interviews to confirm or disconfirm diagnoses.

This review is limited by the K6/K10 translation and cultural adaptation studies identified through the search methods used.

Conclusion

Evidence that the original K6/K10's validity was established for culturally diverse groups is limited, and evidence for the cultural equivalence of translated/adapted K6/K10s in clinical settings is equivocal. The K6/K10's unidimensional structure receives inconsistent support across cultures, as does its predictive validity in primary care, mental health and epidemiological settings. Research on the K6/K10's sensitivity to change with culturally diverse groups is virtually absent. In view of inconsistencies in the evidence for the K6/K10's cultural equivalence, it should be administered in a culturally sensitive manner and scores should not be taken at face value, but should be interpreted with caution in the context of follow-up clinical interviews.

Acknowledgements

This work was commissioned by the Victorian Foundation for Survivors of Torture, with financial support from the Sidney Myer Fund and William Buckland Foundation.

Declaration of interest statement

The authors have no conflicts of interest.

References

Albrecht S.S., McVeigh K.H. (2012) Investigation of the disparity between New York City and national prevalence of nonspecific psychological distress among Hispanics. *Preventing Chronic Disease*, **9**(11_0104), 1–10, DOI: 10.5888/pcd9.110104

Aldworth J., Colpe L.J., Gfroerer J.C., Novak S.P., Chromy J.R., Barker P.R., Barnett-Walker K., Karg R.S., Morton K.B., Spagnola K. (2010) The National Survey on Drug Use and Health Mental Health Surveillance Study: calibration analysis. *International Journal of Methods in Psychiatric Research*, **19**(Suppl 1), 61–87, DOI: 10.1002/mpr.312

Allen N.B., Jackson H.J. (2011) What kind of evidence do we need for evidence-based mental health policy: the case of the Better Access initiative. *Australian and New Zealand Journal of Psychiatry*, **45**(9), 696–699, DOI: 10.3109/00048674.2011.607132

American Educational Research Association (AERA), American Psychological Association, National Council on Measurement in Education. (1999) *Standards for Educational and Psychological Testing*, Washington, DC, American Psychological Association.

Andersen L.S., Grimsrud A., Myer L., Williams D. R., Stein D.J., Seedat S. (2011) The psychometric properties of the K10 and K6 scales in screening for mood and anxiety disorders in the South African Stress and Health study. *International Journal of Methods in Psychiatric Research*, **20**(4), 215–223, DOI: 10.1002/mpr.351

Andrews G., Slade T. (2001) Interpreting scores on the Kessler Psychological Distress Scale (K10). *Australian and New Zealand Journal of Public Health*, **25**(6), 494–497.

Arnaud B., Malet L., Teissedre F., Izaute M., Moustafa F., Geneste J., Schmidt J., Llorca P.M., Brousse G. (2010) Validity study of Kessler's psychological distress scales conducted among patients admitted to French emergency department for alcohol consumption-related disorders. *Alcoholism-Clinical and Experimental Research*, **34**(7), 1235–1245, DOI: 10.1111/j.1530-0277.2010.01201.x

Arnold B.R., Matus Y.E. (2000) Test translation and cultural equivalence methodologies for use with diverse populations. In Cuellar I., Paniagua F.A. (eds) *Handbook of Multicultural Mental Health: Assessment and Treatment of Diverse Populations*, pp 121–136, San Diego, CA, Academic Press.

Australian Institute of Health and Welfare (AIHW). (2009) *Measuring the Social and Emotional Wellbeing of Aboriginal and Torres Strait Islander Peoples*, Cat. no. IHW 24, Canberra, AIHW.

Baggaley R.F., Ganaba R., Filippi V., Kere M., Marshall T., Sombié I., Storeng K.T., Patel V. (2007) Short communication: Detecting depression after pregnancy: the validity of the K10 and K6 in Burkina Faso. *Tropical Medicine and International Health*, **12**(10), 1225–1229, DOI: 10.1111/j.1365-3156.2007.01906.x

Boufous S., Silove D., Bauman A., Steel Z. (2005) Disability and health service utilization associated with psychological distress: the influence of ethnicity. *Mental Health Services Research*, **7**(3), 171–179, DOI: 10.1007/s11020-005-5785-2

Brooks R.T., Beard J., Steel Z. (2006) Factor structure and interpretation of the K10. *Psychological Assessment*, **18**(1), 62–70, DOI: 10.1037/1040-3590.18.1.62

Browne M.A.O., Wells J.E., Scott K.M., McGee M.A. (2010) The Kessler Psychological Distress Scale in Te Rau Hinengaro: the New Zealand Mental Health Survey. *Australian and New Zealand Journal of Psychiatry*, **44**(4), 314–322.

Carra G., Sciarini P., Segagni-Lusignani G., Clerici M., Montomoli C., Kessler R.C. (2011) Do they actually work across borders? Evaluation of two measures of psychological distress as screening instruments in a non Anglo-Saxon country. *European Psychiatry*, **26**(2), 122–127, DOI: 10.1016/j.eurpsy.2010.04.008

Chowdhary N., Patel V. (2010) Detection of common mental disorder and alcohol use disorders in HIV infected people: a validation study in Goa, India. *Asian Journal of Psychiatry*, **3**(3), 130–133, DOI: 10.1016/j.ajp.2010.08.002

Colpe L.J., Barker P.R., Karg R.S., Batts K.R., Morton K.B., Gfroerer J.C., Stolzenberg S.J., Cunningham D.B., First M.B., Aldworth J. (2010) The National Survey on Drug Use and Health Mental Health Surveillance Study: calibration study design and field procedures. *International Journal of Methods in Psychiatric Research*, **19**(Suppl 1), 36–48, DOI: 10.1002/mpr.311

Cunningham J., Paradies Y.C. (2012) Socio-demographic factors and psychological distress in Indigenous and non-Indigenous Australian adults aged 18–64 years: analysis of national survey data. *BMC Public Health*, **12**(95), 1–15, DOI: 10.1186/1471-2458-12-95

Davidson G.R., Murray K.E., Schweitzer R. (2008) Review of refugee mental health and wellbeing: Australian perspectives. *Australian Psychologist*, **43**(3), 16–174, DOI: 10.1080/00050060802163041

Donker T., Comijs H., Cuijpers P., Terluin B., Nolen W., Zitman F., Penninx B. (2010a) The validity of the Dutch K10 and extended K10 screening scales for depressive and anxiety disorders. *Psychiatry Research*, **176**(1), 45–50, DOI: 10.1016/j.psychres.2009.01.012

Donker T., van Straten A., Marks I., Cuijpers P. (2010b) Brief self-rated screening for depression on the Internet. *Journal of Affective Disorders*, **122**(3), 253–259.

Endicott J., Spitzer R.L., Fleiss J.L., Cohen J. (1976) The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disorders. *Archives of General Psychiatry*, **33** (6), 766–771.

Fassaert T., De Wit M.A.S., Tuinebreijer W.C., Wouters H., Verhoeff A.P., Beekman A.T.F., Dekker J. (2009a) Psychometric properties of an interviewer-administered version of the Kessler Psychological Distress scale (K10) among Dutch, Moroccan and Turkish respondents. *International Journal of Methods in Psychiatric Research*, **18**(3), 159–168, DOI: 10.1002/mpr.288

Fassaert T., De Wit M.A.S., Verhoeff A.P., Tuinebreijer W.C., Gorissen W.H.M., Beekman A.T.F., Dekker J. (2009b) Uptake of health services for common mental disorders by first-generation Turkish and Moroccan migrants in the Netherlands. *BMC Public Health*, **9**(307), 1–9, DOI: 10.1186/1471-2458-9-307

Fernandes M.C., Srinivasan K., Stein A.L., Menezes G., Sumithra R.S., Ramchandani P.G. (2011) Assessing prenatal depression in the rural developing world: a comparison of two screening measures. *Archives of Women's Mental Health*, **14**(3), 209–216, DOI: 10.1007/s00737-010-0190-2

First M.B., Gibbon M., Spitzer R.L., Williams J.W. (2002) *Users Guide to the Structured Clinical Interview for DSM-IV-TR Axis I Disorders – Research Version (SCID-I for DSM-IV-TR, November 2002 Revision)*, New York, Biometrics Research Department, New York State Psychiatric Institute.

Furukawa T.A., Kessler R.C., Slade T., Andrews G. (2003) The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental

Health and Well-Being. *Psychological Medicine*, **33**(2), 357–362, DOI: 10.1017/s0033291702006700

Furukawa T.A., Kawakami N., Saitoh M., Ono Y., Nakane Y., Nakamura Y., Tachimori H., Iwata N., Uda H., Nakane H., Watanabe M., Naganuma Y., Hata Y., Kobayashi M., Miyake Y., Takeshima T., Kikkawa T. (2008) The performance of the Japanese version of the K6 and K10 in the World Mental Health Survey Japan. *International Journal of Methods in Psychiatric Research*, **17**(3), 152–158, DOI: 10.1002/mpr.257

Gone J.P., Trimble J.E. (2012) American Indian and Alaska Native mental health: diverse perspectives on enduring disparities. *Annual Review of Clinical Psychology*, **8**(1), 131–160, DOI: 10.1146/annurev-clinpsy-032511-143127

Green S.B., Lissitz R.W., Mulaik S.A. (1977) Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, **37**(4), 827–838.

Grzywacz J.G., Alterman T., Muntaner C., Gabbard S., Nakamoto J., Carroll D.J. (2009) Measuring job characteristics and mental health among Latino farmworkers: results from cognitive testing. *Journal of Immigrant and Minority Health*, **11**(2), 131–138, DOI: 10.1007/s10903-008-9170-2

Haller D.M., Sanci L.A., Sawyer S.M., Patton G.C. (2009) The identification of young people's emotional distress: a study in primary care. *British Journal of General Practitioners*, **59**(560), e61–e70.

Harkness J., Pennell B.-E., Villar A., Gebler N., Aguilar-Gaxiola S., Bilgen I. (2008) Translation procedures and translation assessment in the World Mental Health Survey Initiative. In Kessler R.C., Ustun T.B. (eds) *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*, pp. 91–113, Cambridge, Cambridge University Press.

Henderson S., Andrews G., Hall W. (2000) Australia's mental health: an overview of the general population survey. *Australian and New Zealand Journal of Psychiatry*, **34**(2), 197–205.

Hickie I.B., Andrews G., Davenport T.A. (2002) Measuring outcomes in patients with depression or anxiety: an essential part of clinical practice. *Medical Journal of Australia*, **177**(4), 205–207.

Hickie I.B., Davenport T.A., Luscombe G.M., Moore M., Griffiths K.M., Christensen H. (2010) Practitioner-supported delivery of internet-based cognitive behaviour therapy: evaluation of the feasibility of conducting a cluster randomised trial. *Medical Journal of Australia*, **192**(11), S31.

Hides L., Lubman D.I., Devlin H., Cotton S., Aitken C., Gibbie T., Hellard M. (2007) Reliability and validity of the Kessler 10 and Patient Health Questionnaire among injecting drug users. *Australian and New Zealand Journal of Psychiatry*, **41**(2), 166–168.

Hides L., Elkins K.S., Scaffidi A., Cotton S.M., Carroll S., Lubman D.I. (2011) Does the addition of integrated cognitive behaviour therapy and motivational interviewing improve the outcomes of standard care for young people with comorbid depression and substance misuse? *Medical Journal of Australia*, **195**(3), S31–S37.

Huysamen G.K. (2002) The relevance of the new APA standards for educational and psychological testing for employment testing in South Africa. *South African Journal of Psychology*, **32**(2), 26–33.

Ito M., Nakajima S., Fujisawa D., Miyashita M., Yoshiharu K., Shear M.K., Ghesquiere A., Wall M.M. (2012) Brief measure for screening complicated grief: reliability and discriminant validity. *PLoS ONE*, **7**(2), 1–6, DOI: 10.1371/journal.pone.0031209

Johnson T.P. (1998) Approaches to equivalence in cross-cultural and cross-national survey research. In Harkness J.A. (ed.) *ZUMA-Nachrichten Spezial No. 3. Cross-cultural Survey Equivalence*, pp. 1–40, Mannheim, ZUMA.

Kessler R.C., Andrews G., Colpe L.J., Hiripi E., Mroczek D.K., Normand S.-L.T., Walters E.E., Zaslavsky A.M. (2002) Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, **32**(6), 959–967, DOI: 10.1017/S0033291702006074

Kessler R.C., Barker P.R., Colpe L.J., Epstein J.F., Gfroerer J.C., Hiripi E., Howes M.J., Normand S.-L.T., Manderscheid R.W., Walters E.E., Zaslavsky A.M. (2003) Screening for serious mental Illness in the general population. *Archives of General Psychiatry*, **60**(2), 184–189, DOI: 10.1001/archpsyc.60.2.184

Kessler R.C., Green J.G., Gruber M.J., Sampson N.A., Bromet E., Cuitan M., Furukawa T.A., Gureje O., Hinkov H., Hu C.-Y., Lara C., Lee S., Mneimneh Z., Myer L., Oakley-Browne M., Posada-Villa J., Sagar R., Viana M.C., Zaslavsky A.M. (2010) Screening for serious mental illness in the general population with the K6 screening scale: results from the WHO World Mental Health (WMH) survey initiative. *International Journal of Methods in Psychiatric Research*, **19**(Suppl 1), 4–22.

Kim G., Aguado Loi C.X., Chiriboga D.A., Jang Y., Parmelee P., Allen R.S. (2011) Limited English proficiency as a barrier to mental health service use: a study of Latino and Asian immigrants with psychiatric disorders. *Journal of Psychiatric Research*, **45**(1), 104–110, DOI: 10.1016/j.jpsychires.2010.04.031

Kirmayer L.J., Narasiah L., Munoz M., Rashid M., Ryder A.G., Guzder J., Hassan G., Rousseau C., Pottie K. (2011) Common mental health problems in immigrants and refugees: general approach in primary care. *Canadian Medical Association Journal*, **183**(12), E959–E967, DOI: 10.1503/cmaj.090292

Lalkhen A.G., McCluskey A. (2008) Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care & Pain*, **8**(6), 221–223, DOI: 10.1093/bjaceaccp/mkn041

Laube R.E. (2010) *Symptoms of psychological distress in culturally and linguistically diverse immigrants to Australia* (Doctoral dissertation), Sydney, Australia, Macquarie University. http://macquarie-primo.hosted.exlibrisgroup.com/primo_library/libweb/action/dlDisplay.do?vid=MQ&afterPDS=true&institution=MQ&docId=-MQ_VOYAGER1588523

Lee S., Tsang A., Ng K.L., Ma Y.L., Guo W., Mak A., Kwok K. (2012) Performance of the 6-item Kessler scale for measuring serious mental illness in Hong Kong. *Comprehensive Psychiatry*, **53**(5), 584–592, DOI: 10.1016/j.comppsych.2011.10.001

Loutan L., Bollini P., Pampallona S., Haan D.B.D., Gariazzo F. (1999) Impact of trauma and torture on asylum-seekers. *European Journal of Public Health*, **9**(2), 93–96, DOI: 10.1093/eurpub/9.2.93

Mathieson F., Mihaere K., Collings S., Dowell A., Stanley J. (2012) Maori cultural adaptation of a brief mental health intervention in primary care. *Journal of Primary Health Care*, **4**(3), 231–238.

Mereish E.H., Liu M.M., Helms J.E. (2012) Effects of discrimination on Chinese, Pilipino, and Vietnamese Americans' mental and physical health. *Asian American Journal of Psychology*, **3**(2), 91–103, DOI: 10.1037/a0025876

Mitchell C.M., Beals J. (2011) The utility of the Kessler Screening Scale for Psychological Distress (K6) in two American Indian communities. *Psychological Assessment*, **23**(3), 752–761, DOI: 10.1037/a0023288

Moen P., Kelly E.L., Tranby E., Huang Q. (2011) Changing work, changing health: Can real work-time flexibility promote health behaviors and well-being? *Journal of Health and Social Behavior*, **52**(4), 404–429, DOI: 10.1177/0022146511418979

Moreau N., Hassan G., Rousseau C., Chenguiti K. (2009) Perception that "Everything requires a lot of effort": transcultural SCL-25 item validation. *Journal of Nervous and Mental Disease*, **197**(9), 695–699, DOI: 10.1097/NMD.0b013e3181b3af0c

Morgan C., Mallett R., Hutchinson G., Leff J. (2004) Negative pathways to psychiatric care and ethnicity: the bridge between social science and psychiatry. *Social Science & Medicine*, **58**(4), 739–752, DOI: 10.1016/S0277-9536(03)00233-8

Nagel T., Robinson G., Condon J., Trauer T. (2009) Approach to treatment of mental illness and substance dependence in remote Indigenous communities: results of a mixed methods study. *Australian Journal of Rural Health*, **17**(4), 174–182, DOI: 10.1111/j.1440-1584.2009.01060.x

National Comorbidity Survey (NCS). (2012) K10 and K6 scales. http://www.hcp.med.harvard.edu/ncs/k6_scales.php [5 November 2012].

Nell V. (2000) *Cross-cultural Neuropsychological Assessment: Theory and Practice*, Mahwah, NJ, Lawrence Erlbaum.

Novak S.P., Colpe L.J., Barker P.R., Gfroerer J.C. (2010) Development of a brief mental health impairment scale using a nationally representative sample in the USA. *International Journal of Methods in Psychiatric Research*, **19**(Suppl 1), 49–60, DOI: 10.1002/mpr.313

O'Connor S.S., Beebe T.J., Lineberry T.W., Jobes D.A., Conrad A.K. (2012) The association between the Kessler 10 and suicidality: a cross-sectional analysis. *Comprehensive Psychiatry*, **53**(1), 48–53, DOI: 10.1016/j.comppsych.2011.02.006

Office of the Surgeon General (US). (2001) *Mental Health: Culture, Race, and Ethnicity – A Supplement to Mental Health: A Report of the Surgeon General*, Rockville, MD, US Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services.

Okawa J. (2008) Considerations for the cross-cultural evaluation of refugees and asylum seekers. In Suzuki L.A., Ponterotto J.G. (eds) *Handbook of Multicultural Assessment: Clinical, Psychological and Educational Applications*, pp. 165–194, San Francisco, CA, John Wiley.

Patel V., Araya R., Chowdhary N., King M., Kirkwood B., Nayak S., Simon G., Weiss H.A. (2008) Detecting common mental disorders in primary care in India: a comparison of five screening questionnaires. *Psychological Medicine*, **38**(2), 221–228, DOI: 10.1017/S0033291707002334

Perini S.J., Slade T., Andrews G. (2006) Generic effectiveness measures: sensitivity to symptom change in anxiety disorders. *Journal of Affective Disorders*, **90**(2–3), 123–130, DOI: 10.1016/j.jad.2005.10.011

Pirkis J., Ftanou M., Williamson M., Machlin A., Warr D., Christo J., Castan L., Spittal M.J., Bassilios B., Harris M. (2010) Evaluation of the Better Access to Psychiatrists, Psychologists and GPs through the Medicare Benefits Schedule Initiative. Component A: A Study of Consumers and their Outcomes. Final Report, Melbourne, Centre for Health Policy, Programs and Economics, University of Melbourne. http://www.health.gov.au/internet/main/publishing.nsf/content/6E6AF89CC56D0910CA257848000096DE/$File/A.pdf [16 November 2011].

Prochaska J.J., Sung H.-Y., Max W., Shi Y., Ong M. (2012) Validity study of the K6 scale as a measure of moderate mental distress based on mental health treatment need and utilization. *International Journal of Methods in Psychiatric Research*, **21**(2), 88–97, DOI: 10.1002/mpr.1349

Public Health Division. (2000) Report on the 1997 and 1998 NSW Health Surveys, Sydney, NSW Health Department. http://www0.health.nsw.gov.au/PublicHealth/surveys/hsa/9798/hsindex.htm#mhealth_intro.htm [27 October 2012].

Robins L.N., Wing J., Wittchen H.U., Helzer J.E., Babor T.F., Burke J., Farmer A., Jablenski A., Pickens R., Regier D.A. (1988) The Composite International Diagnostic Interview. An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry*, **45**(12), 1069–1077.

Sakurai K., Nishi A., Kondo K., Yanagida K., Kawakami N. (2011) Screening performance of K6/K10 and other screening instruments for mood and anxiety disorders in Japan. *Psychiatry and Clinical Neurosciences*, **65**(5), 434–441, DOI: 10.1111/j.1440-1819.2011.02236.x.

Sheehan D.V., Lecrubier Y., Sheehan K.H., Amorim P., Janavs J. (1998) The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, **59**(Suppl 20), 22–33.

Shin H.B., Kominski R.A. (2010) Language Use in the United States: 2007, American Community Survey Reports, ACS-12, Washington, DC, US Census Bureau. http://www.census.gov/prod/2010pubs/acs-12.pdf [2 February 2013].

Silove D., Bateman C.R., Brooks R.T., Fonseca C.A. Z., Steel Z., Rodger J., Soosay I., Fox G., Patel V., Bauman A. (2008) Estimating clinically relevant mental disorders in a rural and an urban setting in postconflict Timor Leste. *Archives of General Psychiatry*, **65**(10), 1205–1212, DOI: 10.1001/archpsyc.65.10.1205

Sireci S.G., Parker P. (2006) Validity on trial: psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice*, **25**(3), 27–34.

Slade T., Grove R., Burgess P. (2011) Kessler Psychological Distress Scale: normative data from the 2007 Australian National Survey of Mental Health and Wellbeing. *Australian and New Zealand Journal of Psychiatry*, **45**(4), 308–316, DOI: 10.3109/00048674.2010.543653

Spies G., Stein D.J., Roos A., Faure S.C., Mostert J., Seedat S., Vythilingum B. (2009) Validity of the Kessler 10 (K-10) in detecting DSM-IV defined mood and anxiety disorders among pregnant women. *Archives of Women's Mental Health*, **12**(2), 69–74, DOI: 10.1007/s00737-009-0050-0

Stallman H.M., McDermott B.M., Beckmann M. M., Kay Wilson M., Adam K. (2010) Women who miscarry: the effectiveness and clinical utility of the Kessler 10 questionnaire in screening for ongoing psychological distress. *Australian & New Zealand Journal of Obstetrics & Gynaecology*, **50**(1), 70–76, DOI: 10.1111/j.1479-828X.2009.01110.x

Stolk Y., Minas I.H., Klimidis S. (2008) *Access to Mental Health Services in Victoria: A Focus on Ethnic Communities*, Melbourne, Victorian Transcultural Psychiatry Unit.

Sulaiman-Hill C.M.R., Thompson S.C. (2010) Selecting instruments for assessing psychological wellbeing in Afghan and Kurdish refugee groups. *BMC Research Notes*, **3**, 237–245, DOI: 10.1186/1756-0500-3-237

Sunderland M., Slade T., Stewart G., Andrews G. (2011) Estimating the prevalence of DSM-IV mental illness in the Australian general population using the Kessler Psychological Distress Scale. *Australian and New Zealand Journal of Psychiatry*, **45**(10), 880–889, DOI: 10.3109/00048674.2011.606785

Sunderland M., Mahoney A., Andrews G. (2012a) Investigating the factor structure of the Kessler Psychological Distress Scale in community and clinical samples of the Australian population. *Journal of Psychopathology and Behavioral Assessment*, **34**(2), 253–259, DOI: 10.1007/s10862-012-9276-7

Sunderland M., Hobbs M.J., Anderson T.M., Andrews G. (2012b) Psychological distress across the lifespan: examining age-related item bias in the Kessler 6 Psychological Distress Scale. *International Psychogeriatrics*, **24**(2), 231–242, DOI: 10.1017/S1041610211001852

Tesfaye M., Hanlon C., Wondimagegn D., Alem A. (2010) Detecting postnatal common mental disorders in Addis Ababa, Ethiopia: validation of the Edinburgh Postnatal Depression Scale and Kessler scales. *Journal of Affective Disorders*, **122**(1-2), 102–108, DOI: 10.1016/j.jad.2009.06.020

Transcultural Mental Health Centre (TMHC). (2012) Translated resources: Kessler-10 @(K-10). http://www.dhi.health.nsw.gov.au/default.aspx?ArticleID=536#Kessler [10 October 2012].

Trauer T. (2010) Outcome measurement in chronic mental illness. *International Review of Psychiatry*, **22**(2), 99–113, DOI: 10.3109/09540261003667525

Trauer T. (2011) The public reporting of organizational performance in mental health: coming soon to a mental health service near you. *Australian and New Zealand Journal of Psychiatry*, **45**(6), 432–443, DOI: 10.3109/00048674.2011.566546

UNESCO World Report. (2009) *Investing in Cultural Diversity and Intercultural Dialogue: Executive Summary*. Paris, UNESCO.

Van de Vijver F.J.R. (1998) Towards a theory of bias and equivalence. In Harkness J.A. (ed.) *ZUMA-Nachrichten Spezial No. 3. Cross-Cultural Survey Equivalence*, pp. 41–65, Mannheim, ZUMA. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten_spezial/znspezial3.pdf [5 September 2012].

Van de Vijver F.J.R., Leung K. (2011) Equivalence and bias: a review of concepts, models, and data analytic procedures. In Matsumoto D., Van de Vijver F.J.R. (eds) *Cross-cultural Research Methods in Psychology*, pp. 17–45, Cambridge, Cambridge University Press.

Van de Vijver F.J.R., Poortinga Y.H. (2005) Conceptual and methodological issues in adapting tests. In Hambleton R.K., Merenda P.F., Spielberger C.D. (eds) *Adapting Educational and Psychological Tests for Cross-cultural Assessment*, pp. 39–63, Mahwah, NJ, Lawrence Erlbaum.

Vos T., Barker B., Begg S., Stanley L., Lopez A.D. (2009) Burden of disease and injury in Aboriginal and Torres Strait Islander Peoples: the Indigenous health gap. *International Journal of Epidemiology*, **38**(2), 470–477, DOI: 10.1093/ije/dyn240

Wells J.E., Oakley Browne M.A., Scott K.M., McGee M.A., Baxter J., Kokaua J. (2006) Te Rau Hinengaro: the New Zealand Mental Health Survey: overview of methods and findings. *Australian and New Zealand Journal of Psychiatry*, **40**(10), 835–844.

World Health Organization (WHO). (2012) Management of substance abuse: Process of translation and adaptation of instruments. http://www.who.int/substance_abuse/research_tools/translation/en/ [28 October 2012].

World Health Organization (WHO) World Mental Health Survey (WMHS) Consortium. (2004) Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *JAMA: Journal of the American Medical Association*, **291**(21), 2581–2590, DOI: 10.1001/jama.291.21.2581

Zhu W., Zeng N., Wang N. (2010) Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. Paper presented at the NESUG (North East SAS Users Group): Health Care and Life Sciences Conference, Baltimore, Maryland. http://www.cpdm.ufpr.br/documentos/ROC.pdf [11 February 2013].