HHS Public Access

Author manuscript

Int J Psychol. Author manuscript; available in PMC 2020 June 29.

Published in final edited form as:

Int J Psychol. 2020 June ; 55(3): 425–434. doi:10.1002/ijop.12604.

Validating mental health assessment in Kenya using an innovative gold standard

Leah K. Watson[a,b], **Bonnie N. Kaiser**[b,c], **Ali M. Giusto**[b,d], **David Ayuku**[e], **Eve S. Puffer**[b,d]

[a.]Centre for Global Mental Health, Toronto, Canada

[b.]Duke Global Health Institute, Durham, NC

[c.]Department of Anthropology, Global Health Program, University of California San Diego, La Jolla, CA

[d.]Department of Psychology and Neuroscience, Duke University, Durham, NC

[e.]Moi University, College of Health Sciences, School of Medicine, Department of Behavioral Sciences, Eldoret, Kenya

Abstract

With the growing burden of mental health disorders worldwide, alongside efforts to expand availability of evidence-based interventions, strategies are needed to ensure accurate identification of individuals suffering from mental disorders. Efforts to locally validate mental health assessments are of particular value, yet gold-standard clinical validation is costly, time-intensive, and reliant on available professionals. This study aimed to validate assessment items for mental distress in Kenya, using an innovative gold standard and a combination of culturally adapted and locally developed items. The mixed-methods study drew on surveys and semi-structured interviews, conducted by lay interviewers, with 48 caregivers. Interviews were used to designate mental health “cases” or “non-cases” based on emotional health problems, identified through collaborative clinical rating process with local input. Individual mental health survey items were evaluated for their ability to discriminate between cases and non-cases. Discriminant survey items included 23 items adapted from existing mental health assessment tools, as well as 6 new items developed for the specific cultural context. When items were combined into a scale, results showed good psychometric properties. The use of clinically rated semi-structured interviews provides a promising alternative gold standard that can help address the challenges of conducting diagnostic clinical validation in low-resource settings.

CORRESPONDING AUTHOR: Bonnie N. Kaiser, bfullard@gmail.com.

INDIVIDUAL AUTHOR CONTRIBUTIONS:

EP and DA conceptualized and designed the study. AG co-led initial measures selection, piloting, and adaptation. LW and BK conducted analyses. LW drafted the manuscript, and all other authors edited it.

Compliance with Ethical Standards: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Informed Consent: Informed consent was obtained from all individual adult participants included in the study.

Keywords

Family functioning; validation; measure development; mental health; Kenya

INTRODUCTION

Mental health receives disproportionately little attention in research and policy despite growing recognition of the significant burden of disease caused by mental disorders. In Kenya, mental and substance use disorders are currently the leading cause of years lost due to disability for adults between the ages of 15 and 49 years. Prevalence of major depressive and anxiety disorders particularly contribute to this burden, a trend seen across sub-Saharan Africa (Murray et al., 2012). Major challenges stand in the way of addressing this burden, including a shortage of providers due to limited training and employment opportunities for mental health professionals, varied cultural understandings and expressions of mental health disorders, and lack of universality of diagnostic criteria and screening tools that complicates efforts at identification and linkage to care (Charlson, Diminic, Lund, Degenhardt, & Whiteford, 2014; Stein & Illes, 2015). Despite these challenges, there are growing efforts to implement innovative and cost-effective approaches to delivery of mental healthcare in low-resource settings (Jenkins et al., 2010). A key contributor to the success of these efforts will be availability of locally valid screening tools, which best facilitate identification of individuals in need of care.

The use of locally valid tools is critical for several reasons: it ensures that mental health epidemiology truly reflects the prevalence of mental health problems; it informs intervention development through detection efforts; and it enables appropriate and effective clinical practice, including tracking symptom change over time. In contrast, lack of locally valid measures can result in problems like having to discard findings of epidemiologic surveys intended to inform policy (Gureje, Lasebikan, Kola, & Makanjuola, 2006) and wasting resources through ineffective targeting of interventions (Kohrt, Luitel, Acharya, & Jordans, 2016).

Local validation of tools can begin with assessment tools developed in high-income countries (HICs), often adapted to the local setting. This has been referred to as an etic criterion: use of a specific “outsider” set of criteria as universally valid and reliable (Kaiser, Kohrt, Keys, Khoury, & Brewster, 2013). In Kenya and other low- and middle-income countries (LMICs), studies have largely focused on this etic approach, validating screening tools developed in the US to detect common mental disorders (Adjorlolo & Watt, 2017; Ali, Ryan, & De Silva, 2016). However, studies in African contexts have shown mixed results in terms of validity, highlighting limitations of relying solely on etic criteria (Ali et al., 2016).

As an alternative, several studies have locally developed assessment tools based on culturally and contextually salient ways of experiencing and identifying mental illness (Betancourt et al., 2009; Bolton, 2001; Ice & Yogo, 2005; Murray et al., 2011). This approach, referred to as emic, suggests that assessment of mental health should be based on local illness categories and experiences. The simultaneous use of both etic and emic criteria for identification of mental disorder has also been proposed for mental health assessment in

LMIC settings. Adapted mental health screening tools and/or locally developed tools have been validated and used successfully in multiple settings (Ali et al., 2016; Kaiser et al., 2013; Kohrt et al., 2011; Kohrt et al., 2016; Tsai et al., 2013; Weaver & Kaiser, 2015).

Validation of mental health instruments

A remaining challenge is identification of appropriate gold standards against which to validate mental health assessments. Clinician diagnosis is typically considered the ideal gold standard (Ali et al., 2016). However, due to the severe lack of human resources for mental health in LMICs and the time-intensive nature of clinical interviewing, use of clinician diagnosis as a gold standard is frequently impossible (Tsai et al., 2013).

Validation of mental health assessment tools in cross-cultural contexts therefore remains the subject of extensive debate (Charlson et al., 2014). In sub-Saharan Africa, previous studies have drawn on well-known mental health screening instruments developed in HICs as substitute gold standards. However, poor sensitivity and specificity of both instruments in the setting introduces bias into this validation procedure (Tsai et al., 2013). Others have relied on local informant identification of culturally-specific mental health syndromes as a gold standard (Betancourt et al., 2009; Bolton, 2001; Murray et al., 2011). However, there are yet to be efforts to bridge emic and etic criteria within gold standards, an approach that would be analogous to current recommendations for combining etic and emic content within assessment tools themselves.

This study aimed to validate mental health assessment items using a combination of culturally adapted scales and locally developed items and applying a novel gold standard. We explored which items best discriminated between individuals with or without mental health problems, using semi-structured interviews conducted by non-clinicians as a gold standard for establishing caseness. We also evaluated the performance of the group of resulting items together to examine their potential use as a scale.

METHODS

Setting

This study was part of a broader validity study of measures of family functioning and individual mental health. It was conducted in two peri-urban communities near Eldoret, Kenya. With a population of nearly 300,000, Eldoret is the fifth-largest city in Kenya and is the primary commercial centre in the region. It is home to multiple hospitals and centres of education, including Moi Teaching and Referral Hospital and Moi University. The population of the province is largely of Kalenjin tribal affiliation, but the Luhya, Kikuyu, Kisii, and Luo tribes also have a significant presence.

Recruitment

Participants were recruited via referral by community leaders. Study staff explained the purposes and procedures of the study to the leaders, who were asked to invite families from their community (i.e., within their congregation or neighbourhood) who met these eligibility criteria:

1. At least one caregiver over 18 years old and one child between 8 and 17 years old are currently living in the home.
2. The family is either functioning poorly – i.e., the leader categorizes them as "in need of immediate advising" – or "doing very well."

A Kiswahili consent form was read aloud to participants, who consented via signature or fingerprint. Data collection typically occurred in private locations at local public buildings deemed neutral to staff and participants, such as schools, and was completed by Kenyan research assistants. All procedures were approved by the ethical review boards at Duke University and Moi University School of Medicine.

Surveys

A tablet-based survey was administered to each participant. Table 1 provides an overview of mental health scales considered. Assessment items were selected based on extensive formative work in 2013 with Kenyan mental health professionals, para-professionals, and caregivers. Focus group discussions identified primary domains and specific characteristics in terms of psychological distress and positive mental health. These were used to select relevant assessment tools for inclusion in this study. When qualitative findings yielded important areas that were not reflected in existing measures, new culturally-relevant items were developed. Each item was translated into Kiswahili and back-translated into English. Each item then underwent a cognitive interviewing process with at least three participants, to ensure comprehensibility and acceptability in the setting; this led to clarifications and simplifications, in many cases changing literal translations to ones that better captured items' meanings.

Interviews

Semi-structured interviews assessed emotional and behavioural health, within the context of broader interviews on relational functioning, and lasted approximately 1.5 hours total. The Global Assessment of Relational Functioning (GARF) Scale was used as the model for the overall interview format (Group for the Advancement of Psychiatry Committee on the Family, 1996). We then added domains related to individual caregivers' mental health based on local salience and formative research. Two of these domains – emotional and behavioural health – were used in this study. The emotional health domain captured information about mood and feelings such as stress or "thinking too much." The behavioural health domain captured information about occurrence and frequency of behaviours perceived as negative, including those that interfere with one's relational functioning, such as substance abuse or breaking the law.

The two domains—emotional and behavioural health—were rated on a scale from 1 to 4, with 1 representing severe negative symptoms and 4 representing excellent mental health with no negative symptoms. Mental health case status for each participant was determined based on these scores. Participants rated as a 1 or 2 in the emotional and/or behavioural health domain were designated mental health "cases," and participants rated as a 3 or 4 in the emotional and behavioural health domain were designated "non-cases."

Interview scoring entailed a rigorous, multi-step process involving multiple raters. Audio recordings of interviews were transcribed from Kiswahili directly into English. A team of 4–5 US-based and Kenyan raters independently reviewed transcripts, interviewer notes, transcripts, and audio recordings (Kenyan raters only); all rated participants on all domains. The Kenyan raters included a masters-level clinical psychologist and other individuals with experience in local psychosocial programming but no formal mental health counselling training. The US-based team included a doctoral-level clinical psychologist, a doctoral student in clinical psychology, and trained masters-level research assistants. US and Kenyan teams separately decided upon team consensus ratings, followed by conference calls to determine final ratings. Detailed notes were taken to document the decision-making process, with particular attention to culture- and context-related factors that led to different initial ratings between US and Kenyan rating teams. In cases of difficulty reaching consensus, the Kenyan team ratings were usually given priority.

Analysis

The goal of analysis was to determine which assessment items best discriminated between mental health cases and non-cases, and ultimately to pare down the number of individual mental health items to a measure that is best for screening and monitoring mental health status locally. The validation process used a scoring method similar to that used by Goldberg to develop the General Health Questionnaire (Goldberg, 1978). For each item, the proportion of cases and non-cases highly endorsing the item was calculated (responding 2 or 3 on a scale of 0–3 for all except the PHQ-9, in which case responses of 1–3 were considered high due to different response options; see Appendix). The proportion of non-cases was then subtracted from the proportion of cases to produce a gradient score. A higher gradient score indicates better discrimination; a negative gradient score indicates that a higher proportion of non-cases than cases highly endorsed the item. For example, if 8 out of 10 mental health cases highly endorsed Item A and 2 out of 10 non-cases highly endorsed it, the gradient score would be 0.6. This item would be considered superior to an item with a gradient score of 0.5 in terms of its ability to discriminate between mental health cases and non-cases.

To be considered valid, items had to meet two validity criteria. First, items had to be highly endorsed by fewer than 40% of non-cases (or 40% of cases for positive valence items). Because general distress is endorsed at high levels despite absence of mental disorder in settings like Kenya (Gust et al., 2017), we used a higher level of non-case endorsement than the typical 25%. Second, items needed to have a gradient score greater than 0.25, as lower scores indicated that an item showed poor or no discrimination between cases and non-cases. Exceptions included items that are expected to have low endorsement across the board, such as suicidal ideation – and therefore numerically couldn't achieve a gradient score over 0.25 – and items that formed part of subscales that generally performed well and were therefore all kept. Psychometric performance of the final set of items was evaluated using Cronbach's alpha, area under the Receiver Operating Characteristic curve (AUC), sensitivity, specificity, and diagnostic odds ratio (DOR).

RESULTS

Approximately 60% of female participants were designated mental health cases, whereas only 35% of male participants were (Table 2). 76% of cases qualified based on emotional problems, 8% based on behavioural problems, and 16% based on both emotional and behavioural problems.

A total of 47 candidate survey items were included in the gradient scoring validation procedure. In total, 29 mental health assessment items met validation inclusion criteria (see Table 3). Gradient scores ranged from 0.53 to −0.10 (see Figure 1); the average gradient score across all items was 0.29. Items that met inclusion criteria included all 7 items from the GHQ-D subscale and all 9 items from the PHQ-9 – all designed to measure severe depressive symptomology – as well as all 7 items from the GHQ-B subscale – designed to measure anxiety symptomology (Goldberg, 1978; Spitzer, Kroenke, & Williams, 1999). It should be noted that some of these items were adapted quite significantly while retaining their original meaning. In fact, one of the best-performing items in the sample (GHQA-6) was so heavily adapted that we converted it to being considered a locally developed item (Emic-10), as it was no longer equivalent to the original: “Have you been getting a feeling of tightness or pressure in your head?” was changed to “Do you feel like your head has been pressed like a *chapati*?” (flattened bread). Overall, 6 locally developed items (including Emic-10) met inclusion criteria.

In contrast, most items from the GHQ-A and C subscales – designed to assess somatic symptoms and social dysfunction – failed to meet validation criteria. Overall, the least-discriminant items had a positive valence (e.g., GHQC-6 “have you felt capable of making decisions about things in your life?”), with negative gradient scores indicating that more cases than non-cases highly endorsed them, which is the opposite of expected for positive items.

Psychometric properties of the final measure reflect very high internal consistency (α=0.97) and strong validity in relation to interview-generated caseness categorizations (Table 4). The final scale correctly identified true mental health cases 80% of the time and non-cases 74% of the time. Additionally, a mental health case had 11.3 times greater odds of scoring ≥16 on the final measure than scoring <16.

DISCUSSION

In this study, we aimed to identify mental health assessment items that best discriminate between individuals with and without mental health problems in Kenya. We tested a novel gold standard, first identifying mental health cases and non-cases based on emotional and behavioural health as reported in semi-structured interviews conducted by lay interviewers, and then using a gradient scoring process to compare endorsement of each individual mental health survey item between the mental health cases and non-cases. This yielded a subset of 29 items that met inclusion criteria for discrimination. These items included the full Patient Health Questionnaire 9-item scale (PHQ-9), all 7 adapted items of the General Health Questionnaire (GHQ) subscale measuring depressive symptoms, all 7 adapted items from

the GHQ anxiety subscale, and 6 new items developed for the local context. Psychometric properties suggest that items function as a valid scale.

This study contributes to the need for locally validated tools in the Kenyan setting, for use by both lay people and professionals. We evaluated both locally adapted US screening tools and locally developed items. The qualitative data collection that preceded this study goes beyond the translation and back-translation process found in much of the validation literature of the GHQ and PHQ-9, and it was used to develop novel items as well (Gelaye et al., 2013; Monahan et al., 2009). These approaches improved salience, acceptability, relevance, and comprehensibility of items in the local context, while retaining the meaning of the underlying construct. Results of this study suggest that integration of both culturally adapted and locally developed measures of mental health may provide an advantage over exclusive reliance on measures developed in HICs. The group of items identified in this study allows for comparability across contexts, particularly for instruments like the PHQ-9 that have been validated elsewhere, while also ensuring local relevance of the full set of items using adaptation and item development procedures (Kaiser et al., 2013; Kohrt et al., 2016; Weaver & Kaiser, 2015).

This study also contributes to processes for measure validation in mental health research in low-resource settings. Determination of mental health case status in this study relied upon a novel gold standard, including lay interviewers from the local context and a comprehensive interview rating system applied by multiple raters. Other researchers have utilized modified gold standards for mental health caseness, with some studies using local informants to determine presence of mental illness (Betancourt et al., 2009; Bolton, 2001). This study adds a rigorous interview rating process to determine mental health case status, a process closer to the clinical diagnosis comparison often used in high-resource settings (Bolton, 2001). However, in this resource-constrained context, training lay interviewers in both the administration and rating of interviews proved feasible, and the process of reaching consensus on case identification capitalized on both disciplinary expertise and local experience. Further, the rating procedure ensured that interview content and clinical judgments were based on data and expertise from this setting. At the same time, the rating procedure could easily be applied in a variety of settings. Results of the study demonstrate the utility of this validation method, by which both the gold standard for validation as well as the discriminant screening items were identified in settings lacking mental health professionals, while retaining rigor of the gold standard selection and item assessment processes.

Comparison to validation literature

Our findings that adapted PHQ-9 items perform well in discriminating cases and non-cases are consistent with previous research that used the PHQ-9 for identifying depression in sub-Saharan Africa, including Kenya (Bhana, Rathod, Selohilwe, Kathree, & Petersen, 2015; Cholera et al., 2014; Gelaye et al., 2013; Monahan et al., 2009). However, these validation studies vary with respect to rigor of the choice of gold standard diagnosis, as well as the method of adaptation of items. One study specifically mentioned that a lack of qualitative data informing adaptation of the measure was a limitation to their findings (Cholera et al.,

2014), while others contained little to no detail on adaptation or used only translation and back-translation (Bhana et al., 2015; Gelaye et al., 2013; Monahan et al., 2009). Though fewer studies to-date have validated the GHQ than the PHQ-9 in sub-Saharan Africa, previous literature has shown moderate success of the GHQ in predicting mental health problems in the setting (Abubakar & Fischer, 2012; Gelaye et al., 2015; Makanjuola, Onyeama, Nuhu, Kola, & Gureje, 2014). Similar to the PHQ-9 literature, little cultural adaptation information is given in the studies outside of the translation and back-translation process, suggesting that our adapted items might perform better due to local fit and comprehensibility. Our study addressed some of the methodological limitations of previous work and further confirmed the potential usefulness of these measures overall.

Additionally, six of the eight locally developed items were found to be discriminant. These items addressed general emotional and psychological problems, as well as isolation, frustration, and thoughts. One of these items asked, “Have you had illnesses that relate to a lot of thoughts?” which has been found to be a common idiom of distress cross-culturally (Kaiser et al., 2015). These local items were developed through qualitative research to address salient areas that were not specifically covered in items from the adapted GHQ and PHQ-9 measures.

Likewise, items that did not discriminate between mental health cases and non-cases were consistent with existing measure adaptation and validation literature. First, several items with gradient scores below the 0.25 cutoff had positive valences, meaning that the questions were phrased in a positive manner to the participant (e.g., “Have you been able to enjoy your normal activities?”). Participants highly endorsed these items across both caseness categories, suggesting that they are not locally indicative of distress (or lack thereof). In addition, all four of the worst-performing items had a negative gradient score, indicating that they were highly endorsed by a higher proportion of cases than non-cases despite being conceptually positive and therefore anticipated to be endorsed by more non-cases. This suggests that valence of the item may have influenced how participants responded. This pattern was not found during cognitive interviewing, though questions were asked in isolation during cognitive interviewing without combining positive and negative-valence items. It is therefore possible that their combination within the survey may have resulted in confusion.

This finding corroborates previous research in mental health measure validation, which has shown that phrasing and structure of questions can affect comprehension and ultimately affect answer patterns of respondents, especially when culturally adapting a measure (Kohrt et al., 2011; Kohrt et al., 2016). Specifically, one study conducted in Ghana to test multiple measures of postnatal common mental disorders likewise showed that some positively-worded items did not discriminate between cases and non-cases (Weobong et al., 2009). One helpful lesson from these results is that reverse-scored items may be less beneficial in some contexts, perhaps especially in those where these types of survey measures are less familiar; the disadvantages may outweigh the advantages to this combination of items, which is intended to address concerns about response bias that have been documented in high-resource settings (Kam & Zhou, 2015).

Lastly, some negative valence items also did not discriminate because they were highly endorsed by both individuals with and without emotional or behavioural problems. Wide endorsement of these items regardless of mental health caseness may reflect a generally high level of distress in the setting, captured in items such as “Have you lost sleep because of a lot of worries?” These items therefore may not have differentiated between common stress due to life circumstances and psychopathology, which is the fundamental purpose of screening instruments.

Limitations and Future Directions

Limitations of the study included the small sample size, which limited possible inferences of quantitative analysis, despite otherwise strengthening findings by facilitating collection of multiple data sources for each participant. In addition, the self-report format of the survey and interview introduced the potential for social desirability bias, although this was reduced by using various measures designed to capture one construct in different ways. Relationships between mental health and individual characteristics, such as gender, age, and tribal affiliation could be interesting to explore in future studies. Further, ideally, our innovative gold standard also could be compared with an established gold standard such as the Structured Clinical Interview for DSM-5 (SCID). However, low-resource settings that would most benefit from our approach, such as Kenya, often lack sufficient human resources or validated clinical assessment tools like the SCID. Future research could validate our innovative gold standard in a setting that does have validated standardized tools and more mental health clinicians.

CONCLUSION

Given the increasing burden of mental illness worldwide, there is an urgent need for research regarding mental health screening instruments and methods for validation. This is especially the case in countries that lack mental health professionals and resources. The findings of this study suggest utility of both adapted and locally developed mental health screening items and suggest that integration of both types of items into mental health instruments can provide a more complete picture of how mental health is experienced and expressed locally. Additionally, our validation process is easily replicable and can be used by lay individuals as a gold standard in settings lacking mental health professionals to provide clinical diagnoses. Development of locally salient screening tools should occur in tandem with development of community-based mental healthcare options, or referral to specialists where available.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This research was funded by the Duke Global Health Institute. Dr. Kaiser was supported by the National Institute of Mental Health of the National Institutes of Health under award number F32 MH113288. Drs. Puffer and Kaiser are investigators with the Implementation Research Institute at the George Warren Brown School of Social Work, Washington University in St Louis; through an award from the National Institute of Mental Health (R25 MH08091607) and the Department of Veterans Affairs, Health Services Research & Development Services, Quality Enhancement Research Initiative (QUERI). The authors would like to recognize contributions by Elyse Thulin,

Clare Fisher, Taylor Haynes, Wilter Rono, and the Tuko Pamoja research team. This research was conducted in collaboration with AMPATH.

Funding: This research was funded by the Duke Global Health Institute. Dr. Kaiser was supported by the National Institute of Mental Health of the National Institutes of Health under award number F32 MH113288.

REFERENCES

Abubakar A, & Fischer R (2012). The factor structure of the 12-item General Health Questionnaire in a literate Kenyan population. Stress Health, 28(3), 248–254. doi:10.1002/smi.1420 [PubMed: 22282374]

Adjorlolo S, & Watt BD (2017). Factorial and convergent validity of the Youth Psychopathic Traits Inventory-Short Version in Ghana. Int J Psychol. doi:10.1002/ijop.12468. Advance online publication.

Ali GC, Ryan G, & De Silva MJ (2016). Validated Screening Tools for Common Mental Disorders in Low and Middle Income Countries: A Systematic Review. PLoS One, 11(6), e0156939. doi:10.1371/journal.pone.0156939 [PubMed: 27310297]

Betancourt TS, Bass J, Borisova I, Neugebauer R, Speelman L, Onyango G, & Bolton P (2009). Assessing local instrument reliability and validity: a field-based example from northern Uganda. Soc Psychiatry Psychiatr Epidemiol, 44(8), 685–692. doi:10.1007/s00127-008-0475-1 [PubMed: 19165403]

Bhana A, Rathod SD, Selohilwe O, Kathree T, & Petersen I (2015). The validity of the Patient Health Questionnaire for screening depression in chronic care patients in primary health care in South Africa. BMC Psychiatry, 15, 118. doi:10.1186/s12888-015-0503-0 [PubMed: 26001915]

Bolton P (2001). Cross-cultural validity and reliability testing of a standard psychiatric assessment instrument without a gold standard. J Nerv Ment Dis, 189(4), 238–242. [PubMed: 11339319]

Charlson FJ, Diminic S, Lund C, Degenhardt L, & Whiteford HA (2014). Mental and substance use disorders in Sub-Saharan Africa: predictions of epidemiological changes and mental health workforce requirements for the next 40 years. PLoS One, 9(10), e110208. doi:10.1371/journal.pone.0110208 [PubMed: 25310010]

Cholera R, Gaynes BN, Pence BW, Bassett J, Qangule N, Macphail C, … Miller WC (2014). Validity of the Patient Health Questionnaire-9 to screen for depression in a high-HIV burden primary healthcare clinic in Johannesburg, South Africa. J Affect Disord, 167, 160–166. doi:10.1016/j.jad.2014.06.003 [PubMed: 24972364]

Gelaye B, Tadesse MG, Lohsoonthorn V, Lertmeharit S, Pensuksan WC, Sanchez SE, … Williams MA (2015). Psychometric properties and factor structure of the General Health Questionnaire as a screening tool for anxiety and depressive symptoms in a multi-national study of young adults. J Affect Disord, 187, 197–202. doi:10.1016/j.jad.2015.08.045 [PubMed: 26342172]

Gelaye B, Williams MA, Lemma S, Deyessa N, Bahretibeb Y, Shibre T, … Andrew Zhou XH (2013). Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa. Psychiatry Res, 210(2), 653–661. doi:10.1016/j.psychres.2013.07.015 [PubMed: 23972787]

Goldberg DP (1978). Manual of the general health questionnaire. Windsor, England: NFER Publishing.

Group for the Advancement of Psychiatry Committee on the Family. (1996). Global Assessment of Relational Functioning scale (GARF): I. Background and rationale. Group for the Advancement of Psychiatry Committee on the Family. Fam Process, 35(2), 155–172. [PubMed: 8886770]

Gureje O, Lasebikan VO, Kola L, & Makanjuola VA (2006). Lifetime and 12-month prevalence of mental disorders in the Nigerian Survey of Mental Health and Well-Being. Br J Psychiatry, 188(5), 465–471. doi:10.1192/bjp.188.5.465 [PubMed: 16648534]

Gust DA, Gvetadze R, Furtado M, Makanga M, Akelo V, Ondenge K, … McLellan-Lemal E (2017). Factors associated with psychological distress among young women in Kisumu, Kenya. Int J Womens Health, 9, 255–264. doi:10.2147/IJWH.S125133 [PubMed: 28496366]

Ice GH, & Yogo J (2005). Measuring Stress Among Luo Elders: Development of the Luo Perceived Stress Scale. Field Methods, 17(4), 394–411. doi:10.1177/1525822×05280176

Jenkins R, Baingana F, Belkin G, Borowitz M, Daly A, Francis P, … Sadiq S (2010). Mental health and the development agenda in Sub-Saharan Africa. Psychiatr Serv, 61(3), 229–234. doi:10.1176/ps.2010.61.3.229 [PubMed: 20194398]

Kaiser BN, Haroz EE, Kohrt BA, Bolton PA, Bass JK, & Hinton DE (2015). "Thinking too much": A systematic review of a common idiom of distress. Soc Sci Med, 147, 170–183. doi:10.1016/j.socscimed.2015.10.044 [PubMed: 26584235]

Kaiser BN, Kohrt BA, Keys HM, Khoury NM, & Brewster AR (2013). Strategies for assessing mental health in Haiti: local instrument development and transcultural translation. Transcult Psychiatry, 50(4), 532–558. doi:10.1177/1363461513502697 [PubMed: 24067540]

Kam CCS, & Zhou M (2015). Does Acquiescence Affect Individual Items Consistently? Educ Psychol Meas, 75(5), 764–784. doi:10.1177/0013164414560817 [PubMed: 29795840]

Kohrt BA, Jordans MJ, Tol WA, Luitel NP, Maharjan SM, & Upadhaya N (2011). Validation of cross-cultural child mental health and psychosocial research instruments: adapting the Depression Self-Rating Scale and Child PTSD Symptom Scale in Nepal. BMC Psychiatry, 11(1), 127. doi:10.1186/1471-244X-11-127 [PubMed: 21816045]

Kohrt BA, Luitel NP, Acharya P, & Jordans MJ (2016). Detection of depression in low resource settings: validation of the Patient Health Questionnaire (PHQ-9) and cultural concepts of distress in Nepal. BMC Psychiatry, 16(1), 58. doi:10.1186/s12888-016-0768-y [PubMed: 26951403]

Makanjuola VA, Onyeama M, Nuhu FT, Kola L, & Gureje O (2014). Validation of short screening tools for common mental disorders in Nigerian general practices. Gen Hosp Psychiatry, 36(3), 325–329. doi:10.1016/j.genhosppsych.2013.12.010 [PubMed: 24559789]

Monahan PO, Shacham E, Reece M, Kroenke K, Ong'or WO, Omollo O, … Ojwang C (2009). Validity/reliability of PHQ-9 and PHQ-2 depression scales among adults living with HIV/AIDS in western Kenya. J Gen Intern Med, 24(2), 189–197. doi:10.1007/s11606-008-0846-z [PubMed: 19031037]

Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, … Memish ZA (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet, 380(9859), 2197–2223. doi:10.1016/S0140-6736(12)61689-4 [PubMed: 23245608]

Murray LK, Bass J, Chomba E, Imasiku M, Thea D, Semrau K, … Bolton P (2011). Validation of the UCLA Child Post traumatic stress disorder-reaction index in Zambia. Int J Ment Health Syst, 5(1), 24. doi:10.1186/1752-4458-5-24 [PubMed: 21943178]

Spitzer RL, Kroenke K, & Williams JB (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA, 282(18), 1737–1744. [PubMed: 10568646]

Stein DJ, & Illes J (2015). Beyond Scientism and Skepticism: An Integrative Approach to Global Mental Health. Front Psychiatry, 6, 166. doi:10.3389/fpsyt.2015.00166 [PubMed: 26635641]

Tsai AC, Scott JA, Hung KJ, Zhu JQ, Matthews LT, Psaros C, & Tomlinson M (2013). Reliability and validity of instruments for assessing perinatal depression in African settings: systematic review and meta-analysis. PLoS One, 8(12), e82521. doi:10.1371/journal.pone.0082521 [PubMed: 24340036]

Weaver LJ, & Kaiser BN (2015). Developing and testing locally derived mental health scales: Examples from North India and Haiti. Field Methods, 27(2), 115–130. doi:10.1177/1525822×14547191

Weobong B, Akpalu B, Doku V, Owusu-Agyei S, Hurt L, Kirkwood B, & Prince M (2009). The comparative validity of screening scales for postnatal common mental disorder in Kintampo, Ghana. J Affect Disord, 113(1–2), 109–117. doi:10.1016/j.jad.2008.05.009 [PubMed: 18614241]

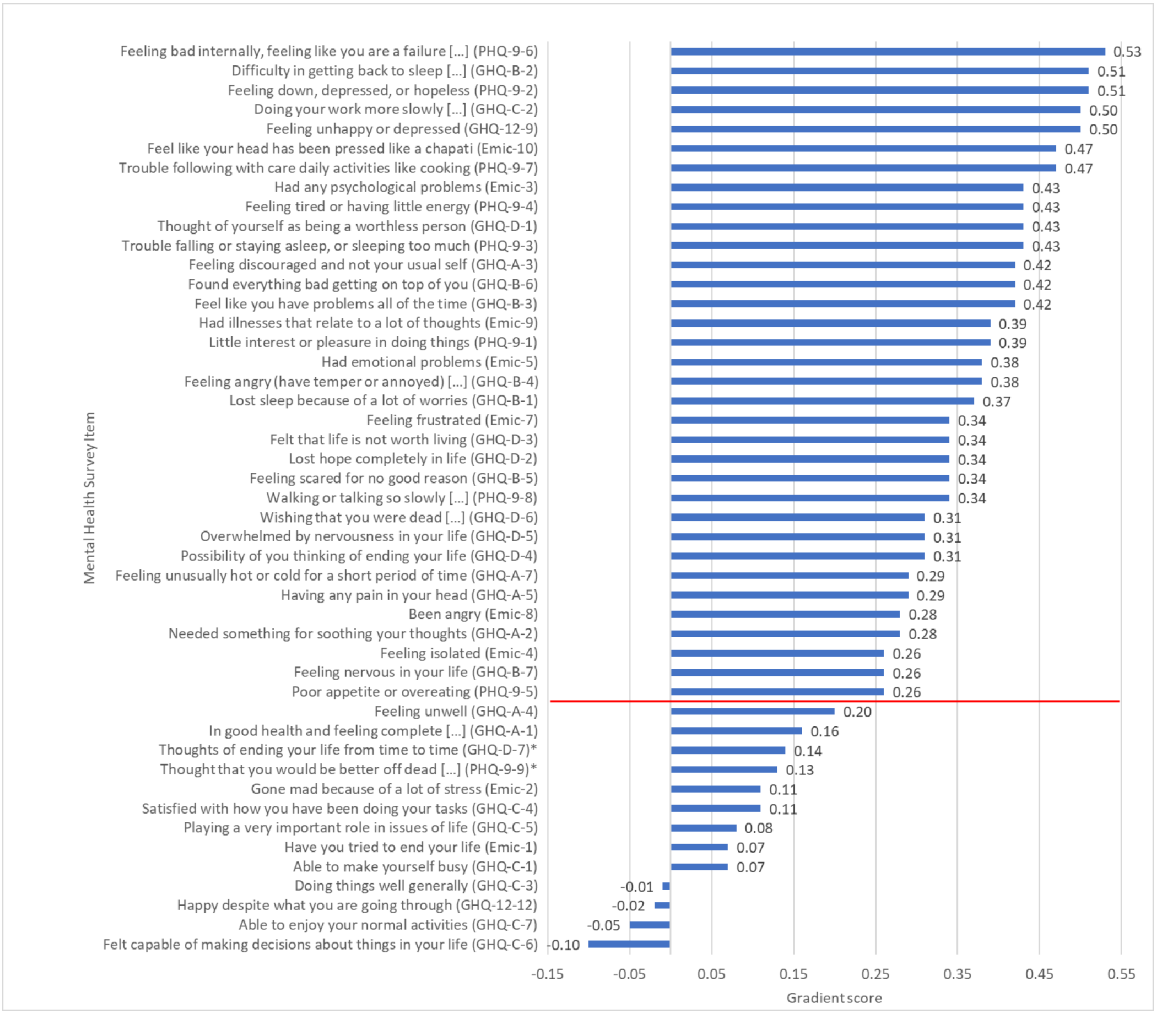

Figure 1:

Mental health survey items by gradient score based on ability to discriminate mental health cases from non-cases (n = 33).

Table 1:

Description of mental health assessment tools and scoring

Scale	Description	Scoring	Number of items
General Health Questionnaire 28-item scale (GHQ-28)	Screens for common minor psychiatric disorders. Contains 4 subscales: somatic symptoms, anxiety and insomnia, social dysfunction, and severe depression.	Likert scale 0–3 based on frequency of symptom, thought, or feeling	28
General Health Questionnaire 12-item scale (GHQ-12)	Screens for common minor psychiatric disorders	Likert scale 0–3 based on frequency of symptom, thought, or feeling	2*
Patient Health Questionnaire 9-item scale (PHQ-9)	Screens for presence and severity of depression	Likert scale 0–3 based on frequency of symptom, thought, or feeling	9
Emic items	Locally-developed items related to adult psychological/emotional health	Likert scale 0–3 based on frequency of symptom, thought, or feeling	8

*Represents items not in the GHQ-28 that were deemed appropriate for the setting based on formative qualitative data collection

Table 2:

Characteristics of study participants

Characteristic: mean (SD[a]) or n (%)	Male (n=17)	Female (n=31)	t-test/chi-square
Age	45.7 (8.6)	35.2 (7.1)	p<0.001
Married	17 (100.0%)	21 (67.7%)	p<0.01
Mean household size	6.7 (2.2)	6.1 (2.1)	p=0.40
Mental health cases as determined by interview	6 (35.3%)	19 (61.3%)	p=0.18[b]

[a] SD = standard deviation

[b] Fisher's exact test

Table 3:

Mental health survey items meeting validation criteria based on gradient scoring procedure[a]

3a: Items included in final assessment tool

Item	Original English item	Case endorsmt. proportion	Non-case endorsmt. proportion	Item gradient score	Reason for inclusion[b]
GHQ-B-1	Lost sleep over worry	0.76	0.39	0.37	
GHQ-B-2	Had difficulty in staying asleep once you are off	0.64	0.13	0.51	
GHQ-B-3	Feel constantly under strain	0.72	0.30	0.42	
GHQ-B-4	Feeling edgy and bad-tempered	0.60	0.22	0.38	
GHQ-B-5	Been getting scared or panicky for no good reason	0.60	0.26	0.34	
GHQ-B-6	Found everything getting on top of you	0.64	0.22	0.42	
GHQ-B-7	Been feeling nervous and strung-up all the time	0.56	0.30	0.26	
GHQ-D-1	Been thinking of yourself as a worthless person	0.60	0.17	0.43	
GHQ-D-2	Felt that life was entirely hopeless	0.56	0.22	0.34	
GHQ-D-3	Felt that life isn't worth living	0.56	0.22	0.34	
GHQ-D-4	Thought of the possibility that you might make away with yourself	0.40	0.09	0.31	
GHQ-D-5	Found at times you couldn't do anything because your nerves were too bad	0.48	0.17	0.31	
GHQ-D-6	Found yourself wishing you were dead and away from it all	0.44	0.13	0.31	
GHQ-D-7	Found that the idea of taking your own life kept coming into your mind	0.36	0.22	0.14	Expect low endorsement
PHQ-9–1	Little interest or pleasure in doing things	0.67	0.28	0.39	
PHQ-9–2	Feeling down, depressed, or hopeless	0.71	0.20	0.51	
PHQ-9–3	Trouble falling or staying asleep, or sleeping too much	0.67	0.24	0.43	
PHQ-9–4	Feeling tired or having little energy	0.79	0.36	0.43	
PHQ-9–5	Poor appetite or overeating	0.50	0.24	0.26	
PHQ-9–6	Feeling bad internally, feeling like you are a failure or that you have let yourself or your family down	0.73	0.20	0.53	
PHQ-9–7	Trouble following with care daily activities like cooking	0.63	0.16	0.47	
PHQ-9–8	Walking or talking so slowly that other people could have noticed. Or the opposite - moving a lot that you have been walking more than usual?	0.54	0.20	0.34	
PHQ-9–9	Thought that you would be better off dead or of hurting yourself in some way	0.29	0.16	0.13	Expect low endorsement
Emic-3	Have you had any psychological problems?	0.60	0.17	0.43	
Emic-4	Have you been feeling isolated?	0.48	0.22	0.26	
Emic-5	Have you had emotional problems?	0.64	0.26	0.38	
Emic-7	Have you been feeling frustrated?	0.56	0.22	0.34	
Emic-9	Have you had illnesses that relate to a lot of thoughts?	0.48	0.09	0.39	

3a: Items included in final assessment tool

Item	Original English item	Case endorsmt. proportion	Non-case endorsmt. proportion	Item gradient score	Reason for inclusion[b]
Emic-10[c]	Do you feel like your head has been pressed like a chapati?	0.60	0.13	0.47	

3b: Items excluded from final assessment tool

Item	Original English item	Case endorsemt. proportion	Non-case endorsemt. proportion	Item gradient score	Reason for exclusion[a]
GHQ-12–9	Been feeling unhappy or depressed	0.72	0.22	0.50	Redundant with PHQ
GHQ-12–12*	Been feeling reasonably happy, all things considered	0.76	0.74	−0.02	
GHQ-A-1*	Been feeling perfectly well and in good health	0.60	0.76	0.16	
GHQ-A-2	Been feeling in need of a good tonic	0.72	0.44	0.28	
GHQ-A-3	Been feeling run down and out of sorts	0.64	0.22	0.42	Removed full scale
GHQ-A-4	Felt that you are ill	0.76	0.57	0.20	
GHQ-A-5	Been getting any pains in your head	0.64	0.35	0.29	Borderline; removed full scale
GHQ-A-7	Been having hot or cold spells	0.68	0.39	0.29	Borderline; removed full scale
GHQ-C-1*	Been managing to keep yourself busy and occupied	0.76	0.83	0.07	
GHQ-C-2	Been taking longer over the things that you do	0.72	0.22	0.50	Removed full scale
GHQ-C-3*	Felt on the whole you were doing things well	0.88	0.87	−0.01	
GHQ-C-4*	Been satisfied with the way you've carried out your task	0.80	0.91	0.11	
GHQ-C-5*	Felt that you are playing a useful part in things	0.88	0.96	0.08	
GHQ-C-6*	Felt capable of making decisions about things	0.88	0.78	−0.10	
GHQ-C-7*	Been able to enjoy your normal day-to-day activities	0.92	0.87	−0.05	
Emic-1	Have you tried to end your life?	0.24	0.17	0.07	
Emic-2	Have you gone mad because of a lot of stress?	0.28	0.17	0.11	
Emic-8	Have you been angry?	0.76	0.48	0.28	

[a] Gradient score >0.25 and mental health non-case endorsement proportion of <40%

[b] For items not meeting validation criteria

[c] GHQ-A-6 was adapted so heavily that it is now considered an emic item

* Positively-worded items are considered valid if significantly more mental health non-cases endorse them compared to cases. Gradient scores for these items are therefore calculated as non-case minus case endorsement.

[a] For items meeting validation criteria

Table 4:

Psychometric properties of final scale (N=48)

Cronbach's alpha	AUC	Sensitivity*	Specificity*	DOR*
0.97	0.80	0.80	0.74	11.3

* Cut-off score=16