Development and Initial Validation of the Counseling Center Assessment of Psychological Symptoms–34

Benjamin D. Locke[1], Andrew A. McAleavey[1], Yu Zhao[1], Pui-Wa Lei[1], Jeffrey A. Hayes[1], Louis G. Castonguay[1], Hongli Li[2], Robin Tate[1], and Yu-Chu Lin[1]

Abstract

A short version of the Counseling Center Assessment of Psychological Symptoms–62 (CCAPS-62) was created via three studies. The final short version (CCAPS-34), which contains 34 items and 7 subscales, demonstrated good discrimination power, support for the proposed factor structure, strong initial convergent validity, and adequate test–retest stability over 1-week and 2-week intervals.

In the past two decades, college counseling centers in the United States observed a substantial increase in the levels of psychopathology and symptom severity among college students (Benton, Robertson, Tseng, Newton, & Benton, 2003). Researchers showed that an increasing number of students experienced psychological problems related to academic skills, relationships, stress, anxiety, depression, and drug use (Benton et al., 2003; O'Malley, Wheeler, Murphey, O'Connell, & Waldo, 1990). However, despite such changes, counseling centers lacked a low-cost, standardized, efficient, and psychometrically sound instrument assessing a range of mental health symptoms tailored specifically to the needs of college students. The original development of the Counseling Center Assessment of Psychological Symptoms (CCAPS) represents an effort to establish such an instrument.

The history and development of the CCAPS-62 was described in detail elsewhere (Locke et al., 2011). In short, the CCAPS-62 was developed to create a relatively brief multidimensional measure suitable for routine use in college counseling centers with college students. Though there are similar measures already in the literature, existing measures do not meet all these needs of counseling centers. Several of the most popular self-report measures used in counseling and psychotherapy

[1]The Pennsylvania State University, University Park, PA, USA
[2]Georgia State University, Atlanta, GA, USA

Corresponding Author:
Benjamin D. Locke, Center for Counseling and Psychological Services, The Pennsylvania State University, 501 Student Health Center, University Park, PA 16802, USA
Email: bdl10@psu.edu

(e.g., the Symptom Checklist-90–R, Derogatis, 1994; the Behavioral Health Monitor-20, Kopta & Lowry, 2002; the Outcome Questionnaire-45, Lambert et al., 1996; and the Treatment Outcome Package, Kraus, Seligman & Jordan, 2005) were not designed specifically for use in counseling centers and do not assess some common problems in this population (e.g., academic distress). Of the few measures that were designed for use in this setting (e.g., College Adjustment Scales, Anton & Reed, 1991; the K-State Problem Identification Rating Scales, Robertson et al., 2006; and the Millon College Counseling Inventory, Millon, Strack, Millon-Niedbala, & Grossman, 2008), less is known about their ability to reliably and validly assess psychological symptoms across multiple counseling centers as well as their appropriateness to assess change in counseling.

A pilot study, using a 70-item version of the CCAPS (CCAPS-70), generated a data set of 19,247 counseling center clients from 52 colleges and universities in the United States. Using this sample and a rational–empirical test development approach, the 70 original items were winnowed to 62 items covering 8 subscales: the CCAPS-62. The eight subscales measure symptoms and distress related to anxiety, depression, social anxiety, academic distress, eating concerns, family distress, hostility, and substance use. Scores of these subscales demonstrated good to excellent internal consistency and test–retest stability over 1-week and 2-week intervals, and scores for each subscale of the CCAPS-62 demonstrated convergent validity with scores from established measures administered concurrently (Locke et al., 2011). Once the CCAPS-62 was completed, it was implemented in counseling centers and continues to be used in more than 100 counseling centers nationwide (Center for Collegiate Mental Health, 2010).

Subsequent discussions with counseling center leadership at national meetings such as the Association for University and College Counseling Center Directors (AUCCCD) and the Association for the Coordination of Counseling Center Clinical Services (ACCCCS) revealed the need for a shorter version requiring only 2 to 3 minutes to complete (rather than 7–10 minutes for the CCAPS-62). A shorter version, which retained the multidimensional structure of the CCAPS-62, would make repeated measurement possible (to monitor treatment progress and outcomes) while also serving as a brief assessment instrument for counseling centers that are pressed for the time, space, and/or technology required for efficient data collection. Thus, the goal of this test-development process was to reduce the length of the CCAPS-62 while maintaining its essential structure and preserving the construct validity of the longer instrument.

The use of short forms of measurement scales, in both research and clinical practice, became popular in recent years (Leite, Huang, & Marcoulides, 2008). Short forms reduce the burden for both respondents and clinicians/researchers while preserving sound psychometric properties, such as reliability and validity of the scores. Two popular approaches to short-form development (i.e., determining which items to retain) are classical test theory (CTT) and item response theory (IRT). The current study used both CTT and IRT methodologies.

Classical Test Theory
Item Selection

When CTT is used to select items, both item difficulty and discrimination may be considered (e.g., Carr, Moss, & Harry, 2005). Whereas item difficulty is indicated by average item scores across examinees (p value for dichotomously scored items or item mean for polytomously scored items), item discrimination (for both dichotomous and polytomous items) can be indicated by the correlation between an item score and the total score. In CTT, items with high discrimination and medium difficulty are generally preferred, especially for norm-referenced tests. Some researchers also select items based on a reduction in the scale's overall Cronbach's

alpha if an item is removed from the scale (e.g., Swindle, Cameron, & Rosen, 2006). A decrease in the Cronbach's alpha when an item is removed indicates good performance of the item and an increase indicates poor performance.

CTT is widely used for item selection because of its ease in implementation and more relaxed sample size requirement (Lei, Wu, DiPerna, & Morgan, 2009). However, CTT item statistics are often criticized for their sample dependence (Hambleton & Swaminathan, 1985). Therefore, short-form development via CTT usually requires a sample that is representative of the intended population (Hambleton & Swaminathan, 1985).

Item Response
Theory Item Selection

In contrast, IRT item statistics are often said to be sample-free if assumptions of the chosen IRT models are satisfied. That is, estimation of the item parameters is independent of the particular sample of examinees drawn from the intended population (Hambleton, 1993, p. 160). In practice, to achieve the sample-free property of IRT models, a large and heterogeneous group of examinees is needed for item parameter calibration (Hambleton & Swaminathan, 1985).

Item selection using the IRT approach is based on estimated item information. The information function in IRT is analogous to reliability in CTT. It is functionally related to the difficulty and discrimination parameters of the items and is inversely related to the error variance of trait estimates ($1/SE^2$). Different from the Reliability Index in CTT, the information function in IRT is conditional on individual levels of the latent trait. The amount of information offered by an item (and subsequently a test) may vary across trait levels. Items that yield the most information at certain trait levels of interest are preferred in item selection. In cases where a single index of reliability is needed to summarize the precision of the entire test, marginal reliability can be calculated (Green, Bock, Humphreys,

Linn, & Reckase, 1984). Marginal reliability is simply 1 minus the ratio of average individual error variances (weighted by the assumed latent trait frequency distribution) to the variance of latent trait scores.

In the current study, the IRT approach was used as the primary method of item selection whereas CTT item statistics were used to facilitate initial item selection and evaluate the results of the IRT approach. As suggested by Lei et al. (2009), use of both CTT and IRT approaches "allow(s) for a more rigorous evaluation of the resulting test forms' psychometric properties, which in turn allowed . . . (for) a more resource-efficient set of measures" (p. 828).

Method

We conducted three studies to develop and assess the short form of the CCAPS (CCAPS-34). The first study reflects a rational–empirical approach to building a short form of an instrument, in which CTT and IRT methodologies were combined with clinical expertise, to produce the CCAPS-34. The second study was an initial attempt to evaluate the proposed factor structure and concurrent validity of the subscale scores of the CCAPS-34 on a sample of undergraduate college students, which is one of the populations of interest for this instrument. The final study examined the test–retest stability of the CCAPS-34 scores over periods of 1 and 2 weeks. Each study is described and discussed separately prior to a general discussion.

Study 1

Sample and Instrument

Clients ($N = 19,247$) who sought counseling at 52 college counseling centers across the United States during the Fall 2008 semester completed the 70-item CCAPS. Participants aged from 18 to 63 years ($M = 22.6$, $SD = 5.07$). The sample comprised more females (64%) than males (35%) and more undergraduate students (82.7%) than graduate students

(14.9%). The majority of the sample (70.3%) identified themselves as White/Caucasian. Listwise deletion was used to remove cases with missing data because the sample size was large and the proportion of missing cases was small. A total of 19,082 participants provided complete data for the current study, and this sample was described elsewhere (Locke et al., 2011). Because the sample was reasonably large, we randomly split it into two data sets: one ($n = 9576$) was used for parameter estimation and item selection, and the other ($n = 9,506$) for cross-validation (Locke et al., 2011).

The items in the CCAPS-70 cover a broad spectrum of psychological symptoms. They describe thoughts, feelings, and experiences that people may have in everyday life. Participants were asked to indicate how well each item in the CCAPS described them in the past 2 weeks by rating on a 5-point Likert-type scale ranging from 0 (*not at all like me*) to 4 (*extremely like me*). Ten items were reverse-worded to check the consistency of responses, and they were reverse-coded so that higher scores reflect higher distress.

Procedure

Factor analyses on the CCAPS-70 items were conducted, and results were reported in Locke et al. (2011). Sixty-two items, representing eight factors, were retained, and the factors were labeled Depression, Generalized Anxiety, Social Anxiety, Academic Distress, Eating Concerns, Hostility, Substance Use, and Family Distress. This 62-item, 8-factor instrument is the CCAPS-62. For the purpose of developing a short form for the CCAPS, item selection was done separately for each factor or subscale of the CCAPS-62 to meet the IRT unidimensionality assumption.

Classical item analyses were first conducted separately for each factor of the CCAPS-62. All items that loaded on a factor of the CCAPS-62 were included in a subscale. For each subscale, mean item responses, item-total correlation, and Cronbach's alpha without the item were examined to provide an

initial sense of the quality of individual items on the subscale. Cronbach's alpha was calculated for each subscale that served as a criterion for the shortened version.

The graded response IRT model (GRM, Samejima, 1969) was used to estimate item parameters and information because the CCAPS items were polytomously scored. The MULTILOG software program (Thissen, 1991) was used for all IRT model fitting. Estimated item information curves from the GRM were examined for each item. Items that yielded the most information at the trait levels of interest (in this case, the middle to upper range of the latent trait continuum, because the general population of students in counseling centers is of interest) were considered to be the best-performing items for each subscale.

Sets of candidate items were selected for each subscale based on the CTT and IRT item selection results (see next section). The resulting candidate item sets were then carefully reviewed by a group of content experts that consisted of (a) more than 10 senior staff clinicians at a large university counseling center with decades of clinical experience and at least 1 year of experience using the CCAPS in clinical practice and (b) 2 psychotherapy researchers with greater than 20 years each of clinical and research experience. Taking into consideration both the psychometric properties and the content representativeness of the items, the content experts deliberated and selected the sets of items that best represented the content of the latent factors being studied while also offering clinical utility. They also tried to balance the number of items per subscale according to the importance of the subscale to the entire psychological measure. One important choice during this process was to remove the Family Distress subscale because of the perception that it was less likely to change during short-term treatment and did not need to be assessed each week, which is one intended use of a short form. In addition, it was determined that the shortened form of the Substance Use subscale would be more uniform in content if it focused on

only one substance and so was renamed Alcohol Use.

Multiple short versions were considered for each subscale by examining their content, projected test information functions, and projected Cronbach's alphas. Projected Cronbach's alpha estimates were also computed based on the independent validation sample for the various short forms to ensure that the estimates were not biased due to same sample analyses. Ideally, when the final shortened version is compared with its parent instrument, it should not lose much information, have acceptably high Cronbach's alphas for its scores (e.g., >.70 for instrument development as suggested by Henson, 2001), and have good content representation of the long version.

Results

The final short form of the CCAPS contains 34 items spanning seven subscales. Because of space considerations, this section will discuss in detail the results of item selection for the Depression subscale. Item selection for other subscales followed the exact same procedure.

Classical Test Theory Item Analysis

Results of CTT item analysis are shown in Table 1. Cronbach's alpha (.918) and marginal reliability estimate from GRM (.930) for the original 14-item Depression subscale scores were very high. Table 1 shows that in general, all items performed well. The best items were Items 13 (with corrected item–total correlation, $r = .747$ and Cronbach's alpha without the item, $\alpha_{-d} = .908$), 24 ($r = .748$, $\alpha_{-d} = .909$), 27 ($r = .770$, $\alpha_{-d} = .908$), and 45 ($r = .781$, $\alpha_{-d} = .908$). These items had comparatively high discrimination, and if deleted, Cronbach's alpha for the entire subscale would decrease. One item that was not performing well was Item 65. It had relatively low discrimination ($r = .488$), and Cronbach's alpha for the subscale would increase if the item was deleted from the measure.

Item Response Theory Item Analysis

Results of the IRT item information estimation for the Depression subscale are shown in Figure 1. Consistent with the CTT item analysis results, the IRT item information also showed that Items 13, 24, 27, and 45 were the best-performing items on the Depression subscale. These items provided the most information around the middle range of the latent trait (depression). On the other hand, Item 65 was less than ideal and was removed from the short form. It did not give much information across trait levels. This result was also consistent with the CTT item analysis, and the content of the item (which assesses for frequency of crying spells) was not judged to be essential to the clinical relevance of the instrument.

The rest of the items in the subscale performed very similarly in terms of item information. Figure 1 shows that the item information curves for Items 11, 61, and 70 nearly overlapped, as did the item information curves for Items 15, 32, 41, and 55. The nearly overlapping item information curves suggested that the items gave similar amount of information in the similar range of the latent trait. In addition, Items 10 and 51 were also similar in terms of the amount of information they offered; however, the item information curve of Item 10 reached the peak in the middle of the latent trait continuum while that of Item 51 was negatively skewed, suggesting that Item 10 provided the most information in the middle range whereas Item 51 offered more information in the upper range of the latent trait.

Item Selection

Based on the statistical analyses of the items in the Depression subscale, the content experts selected Items 11, 13, 24, 27, 45, and 51 for the short version of the subscale. Items 13, 24, 27, and 45 were the best-performing items as discussed above. Items 11 and 51 did not perform as well as the rest of the items in the short form, but their performance was reasonably

Table 1. Classical Item Statistics

	Item No.	Mean	Standard Deviation	Corrected Item–Total Correlation	Cronbach's α if Item Deleted
Depression	10	2.04	1.379	.687	.911
	11	1.76	1.412	.636	.913
	13	2.00	1.383	.747	.908
	15	1.12	1.240	.568	.915
	24	1.35	1.322	.748	.909
	27	1.65	1.364	.770	.908
	32	1.83	1.130	.581	.914
	41	1.51	1.432	.565	.915
	45	1.64	1.312	.781	.908
	51	0.56	1.028	.536	.916
	55	1.60	1.457	.576	.915
	61	1.64	1.138	.648	.912
	65	1.55	1.494	.488	.919
	70	1.57	1.405	.646	.912
Generalized Anxiety	4	1.64	1.217	.521	.835
	5	1.35	1.343	.620	.824
	17	0.94	1.305	.616	.825
	21	2.00	1.472	.483	.840
	22	2.28	1.403	.577	.829
	31	1.14	1.357	.680	.818
	34	2.32	1.334	.610	.825
	37	1.22	1.303	.517	.835
	44	1.02	1.344	.448	.842
Social Anxiety	3	1.76	1.222	.671	.807
	19	2.07	1.406	.427	.846
	39	1.71	1.191	.556	.824
	46	1.67	1.411	.554	.825
	49	1.74	1.330	.719	.798
	52	1.89	1.377	.679	.805
	60	1.86	1.162	.580	.821
Academic Distress	8	1.70	1.080	.472	.827
	18	1.59	1.214	.575	.802
	57	2.26	1.473	.600	.797
	59	2.18	1.456	.754	.746
	66	1.65	1.408	.709	.762
Eating Concerns	6	0.84	1.184	.778	.879
	16	0.90	1.284	.775	.879
	23	2.05	1.256	.567	.896
	26	1.70	1.435	.690	.887
	29	0.98	1.268	.776	.879
	35	0.65	1.072	.731	.884
	38	0.74	1.162	.606	.892
	53	0.22	0.723	.433	.903
	69	0.79	1.230	.647	.889
Hostility	36	1.17	1.247	.750	.818
	40	1.23	1.327	.632	.837
	48	1.29	1.288	.790	.811

(continued)

Table 1. (continued)

	Item No.	Mean	Standard Deviation	Corrected Item–Total Correlation	Cronbach's α if Item Deleted
	50	1.69	1.317	.590	.844
	58	0.51	0.996	.609	.840
	64	0.92	1.148	.600	.841
	68	0.22	0.656	.421	.863
Substance Use	28	0.34	0.838	.396	.850
	30	0.79	1.110	.756	.785
	33	0.45	0.891	.632	.814
	54	0.56	1.028	.770	.785
	56	1.17	1.299	.651	.808
	63	0.96	1.378	.568	.831

Note: Item numbers are derived from the Counseling Center Assessment of Psychological Symptoms–70 (CCAPS-70). Entries in boldface indicate items that were retained in the final CCAPS-34.

good and their content was deemed to be particularly important to the clinical utility of the measure, particularly Item 51, which assesses for suicidal ideation. Projected Cronbach's alpha for scores on the shortened version is .876, compared with .918 for the scores on the original 14-item long form. The projected GRM marginal reliability estimate for scores is .882, compared with .930 for the original long-form scores. This indicates that the short form for the Depression subscale has adequate psychometric properties in that it does not lose much information compared with the original subscale, and it is expected to yield reasonably high Cronbach's alpha and marginal reliability estimates.

Other Short Forms and Cross-Validation

Following the same procedure, items were selected by psychometricians and content experts jointly for each subscale. Projected Cronbach's alpha and GRM marginal reliability estimates were monitored for each subscale short form. The items that constitute each final short form are highlighted in boldface in Table 1. Cronbach's alphas and GRM marginal reliability estimates for the original and short-form subscale scores based on both the training and cross-validation samples are given in Table 2. Table 2 shows that same sample reliability estimates were similar in magnitude to the cross-validated counterparts estimated using an independent sample. All the final short forms were projected to yield relatively high internal consistency and marginal reliability estimates. In some of the subscales, such as Depression and Eating Concerns, more than half of the items were removed but the projected reliability indices did not decrease substantially. This was especially true for the Eating Concerns subscale. For the short form of this subscale, only Items 6, 16, and 29 were retained because their information functions were significantly higher than the other items across most of the levels of the latent trait, and there was agreement with the content experts that the other items contained considerable overlap. Although the length of the measure went from nine items down to three items, Cronbach's alpha was expected to decrease only by .008 (from .899 for the original form to .891 for the short form).

Discussion

In this study, we developed a short form of a measure of psychological symptoms designed to be used with college students, particularly in counseling. The CCAPS-34
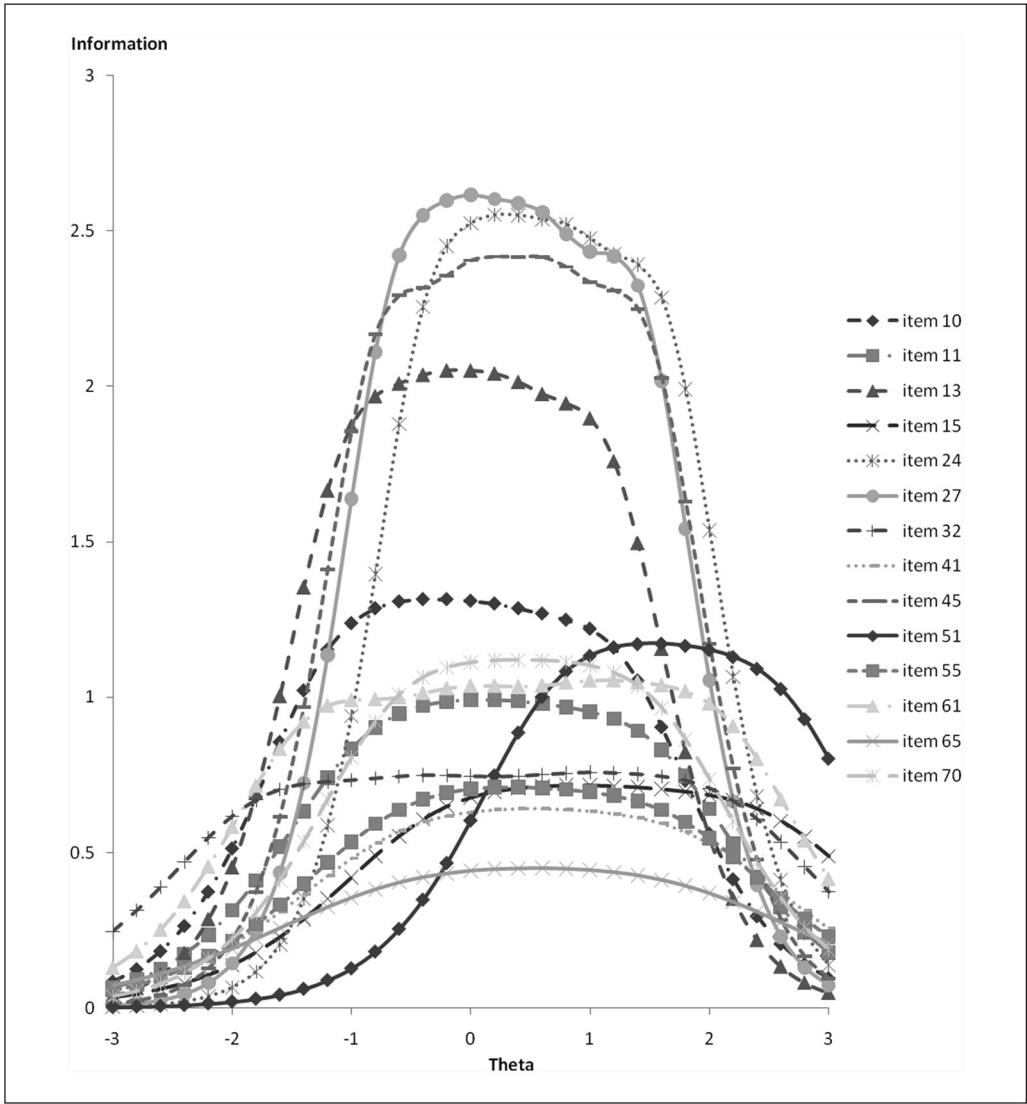
Figure 1. Estimated item information: *Depression* subscale

was developed using input from CTT, IRT, and clinical expertise. The input for these analyses consisted of a large sample of college counseling clients. The emerging measure contains 34 items believed to assess 7 related but distinct types of psychological symptoms most likely to be present in college students when seeking college counseling.

Study 2

Because the performance of items on an instrument can vary somewhat when administered within a short form rather than as part of the original scale, we conducted Study 2 to test the proposed factor structure of the CCAPS-34 to examine the evidence for convergent validity of the CCAPS-34 subscales when

Table 2. Reliability Coefficients for the Original CCAPS-62 and Short-Form CCAPS-34 Subscale Scores

Subscale	No. of Items	Train	Cross	Marginal Reliability Train	Marginal Reliability Cross
Depression					
Original	14	.918	.919	.930	.931
SF	6	.876	.879	.882	.883
Generalized Anxiety					
Original	9	.846	.849	.866	.868
SF	6	.824	.827	.841	.843
Social Anxiety					
Original	7	.840	.839	.873	.873
SF	5	.827	.826	.847	.846
Academic Distress					
Original	5	.824	.824	.858	.860
SF	4	.827	.826	.845	.846
Eating Concerns					
Original	9	.899	.899	.865	.866
SF	3	.891	.893	.726	.727
Hostility					
Original	7	.857	.860	.860	.860
SF	6	.844	.844	.834	.834
Substance/Alcohol Use					
Original	6	.840	.847	.772	.778
SF	4	.823	.831	.704	.715

Note: CCAPS = Counseling Center Assessment of Psychological Symptoms; SF = short form; Train = training subsample; Cross = cross-validation subsample.

administered independent of the larger CCAPS-70 in a sample of college students.

Participants

College undergraduates ($N = 483$) from one large mid-Atlantic university completed this study in Fall 2010 for course credit. Of these, one participant provided no usable data and was excluded from all analyses. Of the remaining, 371 (77%) were female, 403 (84%) were European American, 28 (5.8%) were Asian or Asian American, 17 (3.5%) were Black or African American, and 12 (2.5%) were Hispanic or Latino/a, with remaining ethnic/cultural groups accounting for less than 5% of the sample combined. The vast majority of this sample was heterosexual (95%) and noninternational students (97%). The mean age was 18.85 years ($SD = 1.43$), and 283 were first-year students (59%), 92 were sophomores (19%), 79 were juniors (16%), and 24 (5%) were seniors. In all, 19 (4%) participants reported currently participating in counseling for mental health concerns, and 22 (5%) participants reported taking prescribed psychiatric medications for mental health concerns. Participants reporting current psychological treatment were included in the sample for this study.

For the confirmatory factor analysis (CFA), all 482 participants were included in the analyses because all participants provided complete data on the CCAPS-34. For the correlational analyses of convergent evidence, pairwise deletion was used to limit the effective sample size of correlations to only those participants who answered every item in the relevant scales or subscales. Pairwise deletion was selected because though all participants completed the CCAPS-34, the remaining questionnaires were lengthy (requiring roughly 45 minutes to 1 hour), and most participants failed to complete every item. This led to sample sizes ranging from 413 to 467 for the correlation coefficients.

Measures

To maximize validity of comparisons between the CCAPS-34 and the CCAPS-62, the same referent measures were used in this study as in the study reported by Locke et al. (2011). These measures were selected by those authors as appropriate comparison measures for the constructs on the CCAPS-62.

Alcohol Use Disorders Identification Test. The Alcohol Use Disorders Identification Test (AUDIT; Saunders, Aasland, Babor, de la Fuente, & Grant, 1993) is a 10-item measure of problematic alcohol use that demonstrates adequate evidence of validity (e.g., sensitivity of .86 and specificity of .81 for hazardous consumption of alcohol) and internal consistency ($\alpha = .88$ in an American sample; Saunders et al., 1993). In this sample, the Cronbach's alpha was .780.

Beck Depression Inventory. The Beck Depression Inventory (BDI; Beck, Ward, Mendelsohn, Mock, & Erbaugh, 1961) is a 21-item self-report measure designed to assess symptoms of depression. The BDI demonstrated strong evidence of convergent validity with other measures of depression symptoms (e.g., $r = .72$ with the Hamilton Rating Scale for Depression; Coles, Gibb, & Heimberg, 2001). Internal consistency was reported for scores on the BDI at $\alpha = .86$ on average (Beck, Steer, &

Garbin, 1988). In this sample, the Cronbach's alpha was .877.

Beck Anxiety Inventory. The Beck Anxiety Inventory (BAI; Beck, Epstein, Brown, & Steer, 1988) is a 21-item self-report measure designed to assess symptoms and severity of anxiety. The BAI demonstrated adequate evidence of internal consistency ($\alpha = .92$) and concurrent validity (e.g., $r = .81$ with the SCL-90-R Anxiety scale; Steer, Ranieri, Beck, & Clark, 1993). In this sample the Cronbach's alpha was .923.

Social Phobia Diagnostic Questionnaire. The Social Phobia Diagnostic Questionnaire (SPDQ; Newman, Kachin, Zuellig, Constantino, & Cashman-McGrath, 2003) is a 25-item self-report measure designed to assess the symptoms of social phobia as defined in the *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition (*DSM-IV*). In samples of undergraduate college students, the SPDQ demonstrated evidence of clinical validity in terms of correct diagnostic classification, and correlated highly with other measures of social anxiety (e.g., $r = .61$ with the Social Avoidance and Distress Scale; Newman et al., 2003), with excellent internal consistency ($\alpha = .95$). In this sample, the Cronbach's alpha was .922.

Student Adaptation to College Questionnaire. The Student Adaptation to College Questionnaire (SACQ; Baker & Siryk, 1986) is a 67-item Likert-type self-report measure designed to assess undergraduate students' adjustment to the college environment. The SACQ is scored using four subscales: Academic Adjustment, Social Adjustment, Personal–Emotional Adjustment, and Institutional Attachment, and a total score. Because the SACQ assesses a much broader range of constructs than does the CCAPS Academic Distress scale, the SACQ Academic Adjustment scale (internal consistency for its scores has been reported at $\alpha = .84$; Baker & Siryk, 1986) was used in the analysis. In this sample, the Cronbach's alpha was .874 for the Academic Adjustment subscale.

Eating Attitudes Test. The Eating Attitudes Test (EAT-26; Mintz & O'Halloran, 2000) is

a 26-item measure designed to assess problematic attitudes and behaviors related to eating. The EAT-26 scale demonstrated strong internal consistency ($\alpha = .90$) and correlates with other measures of eating concerns (e.g., $r = .57$ with a body image composite score; Garner, Olmsted, Polivy, & Garfinkel, 1982). In this sample, the Cronbach's alpha was .893.

State-Trait Anger Expression Inventory–2. The State-Trait Anger Expression Inventory–2 (STAXI-2; Spielberger, 1999) is a 57-item inventory designed to measure the experience, expression, and control of anger in adolescents and adults. The STAXI-2 is an expansion of the original STAXI, which demonstrated strong evidence of reliability and validity (Fuqua et al., 1991; Spielberger, 1999). Because the CCAPS Hostility subscale asks respondents to rate themselves over the past 2 weeks, only the Trait Anger subscale of the STAXI-2 is used in the present analysis; in a study of college students, the Trait Anger subscale scores had good internal consistency ($\alpha = .84$; Agliata & Renk, 2009) and have correlated with other measures of hostility (e.g., $r = .66$ with the Buss–Durkee Hostility Inventory; Spielberger, 1999). In this sample, the Cronbach's alpha was .869.

Marlowe–Crowne Social Desirability Scale– Short Version. The Marlowe–Crowne Social Desirability Scale (MCSD; Reynolds, 1982) is a 13-item inventory designed to assess social desirable responding. The MCSD demonstrated good evidence of reliability (Kuder–Richardson coefficient of reliability = .76) and concurrent validity (e.g., $r = .41$ with the Edwards Social Desirability Scale; Reynolds, 1982) and is frequently used in self-report research. In this sample, the Cronbach's alpha was .70.

Procedure

The participants completed all measures online, after being instructed to complete them all in one sitting and completing the informed consent. Participants were randomly assigned to one of three groups. Each group completed a brief demographic questionnaire and the CCAPS-34 prior to the remaining measures, which were given in three counterbalanced orders. These groups were combined for all analyses because they did not differ significantly in a multivariate analysis of variance on the counterbalanced referent measures, $F(18, 916) = 1. 535$, $p = .071$, *ns*.

Results

To test the factor structure of the CCAPS-34 when administered separately from a longer measure, we conducted a CFA in LISREL version 8.80 (Jöreskog & Sörbom, 2006) using robust diagonally weighted least squares estimation. We requested the polychoric and asymptotic covariance matrices using PRELIS (Jöreskog & Sörbom, 2006) and used these as the input. This method was selected to be consistent with Locke et al. (2011) and account for the ordinal nature of the Likert-type responses. Items were allowed to load only on one factor; factor variances were fixed to 1 to scale the factors; factors were permitted to correlate with one another, and no correlated errors were allowed among the items.

The factor structure appeared to fit the data well. While the null hypothesis of exact fit was rejected by the Satorra–Bentler Scaled chi-square test, $\chi^2(506) = 1096.05$, $p < .001$, other measures of global fit indicated good to close model fit to the data. The comparative fit index (CFI) = .98, nonnormed fit index (NNFI) = .98, and incremental fit index (IFI) = .98, were all greater than the .95 cutoff recommended by Hu and Bentler (1999); root mean square error of approximation (RMSEA) = .049 (90% confidence interval = .045, .053), which was just less than the .05 criterion suggested by Browne and Cudeck (1993) for close fit; and the standardized root mean square residual (SRMR) = .063 was less than the .08 cutoff suggested by Hu and Bentler (1999). All factor loadings were significant and generally high, from .47 to .98 across all items (mean = .79). Ranges (denoted within brackets) and means for the subscale factor loadings were Depression [.73, .95],

Table 3. Factor Intercorrelations of the CCAPS-34 in the CFA Model

	Depression	Generalized Anxiety	Social Anxiety	Academic Distress	Eating Concerns	Hostility	Alcohol Use
Depression	1.00						
Generalized Anxiety	0.76*	1.00					
Social Anxiety	0.79*	0.60*	1.00				
Academic Distress	0.68*	0.72*	0.59*	1.00			
Eating Concerns	0.46*	0.45*	0.40*	0.52*	1.00		
Hostility	0.69*	0.65*	0.46*	0.62*	0.40*	1.00	
Alcohol Use	0.23*	0.36*	0.09	0.40*	0.43*	0.35*	1.00

Note: CCAPS = Counseling Center Assessment of Psychological Symptoms; CFA = confirmatory factor analysis. Correlation coefficients between latent factors of the CCAPS-34 in the CFA analysis ($n = 482$).
*$p < .05$.

Table 4. Convergent Validity Evidence of the CCAPS-34 Subscale Scores

CCAPS-34 Subscale	α	M	SD	BDI	BAI	SPDQ	SACQ Academic Adjustment	EAT-26	STAXI-2 Trait Anger	AUDIT	MCSD
Depression	.892	0.70	0.80	.698**	.547**	.569**	−.416**	.357**	.423**	.091*	−.353**
Generalized Anxiety	.820	1.02	0.82	.570**	.678**	.461**	−.341**	.304**	.354**	.161**	−.276**
Social Anxiety	.796	1.53	0.85	.450**	.362**	.760**	−.363**	.196**	.292**	−.047	−.298**
Academic Distress	.760	0.93	0.76	.559**	.492**	.415**	−.681**	.293**	.318**	.222**	−.395**
Eating Concerns	.871	1.15	1.07	.377**	.334**	.267**	−.281**	.520**	.262**	.197**	−.309**
Hostility	.854	0.61	0.68	.471**	.432**	.265**	−.390**	.276**	.691**	.207**	−.440**
Alcohol Use	.869	0.77	0.91	.206**	.247**	.087	−.219**	.236**	.183**	.777**	−.261**
n		482		443	454	442	432	425	459	467	464

Note: CCAPS = Counseling Center Assessment of Psychological Symptoms; AUDIT = Alcohol Use Disorders Identification Test; BAI = Beck Anxiety Inventory; BDI = Beck Depression Inventory; EAT-26 = Eating Attitudes Test–26; MCSD = Marlowe–Crowne Social Desirability Scale; SACQ = Student Adaptation to College Questionnaire; SPDQ = Social Phobia Diagnostic Questionnaire.
*$p < .05$. **$p < .01$.

mean = .85; Generalized Anxiety [.62, .84], mean = .74; Social Anxiety [.47, .89], mean =.68; Academic Distress [.47, .90], mean =.73; Eating Concerns [.85, .89], mean =.87; Hostility [.73, .91], mean =.84; Alcohol Use [.82, .98], mean =.88. Factor intercorrelations can be found in Table 3.

Subscale scores of the CCAPS-34 showed acceptable to good internal consistency in this sample (see Table 4). Cronbach's alpha

coefficients ranged from .760 (Academic Distress) to .892 (Depression). It should be noted that these alpha coefficients are largely comparable to those projected in the development process (Study 1, see Table 2).

Correlation coefficients between CCAPS-34 subscale scores and individual referent measures are reported in Table 4. Each subscale of the CCAPS-34 correlated most strongly with its hypothesized referent measure. These

Pearson product–moment correlations ranged from $r = .520$ (for the Eating Concerns subscale with the EAT-26 total score) to $r = .777$ (for the Alcohol Use subscale with the AUDIT). These convergent correlations are all above Cohen's (1988) suggested criterion of $r = .5$ for a "large" effect. The negative correlations between CCAPS-34 subscale scores and the MCSD are consistent with those reported by Locke et al. (2011).

Discussion

Overall, the results of Study 2 suggest that the CCAPS-34 conforms reasonably well to the factor structure that was presumed during development and based on the CCAPS-62. In addition, the subscale scores of the CCAPS-34 showed evidence of convergent validity in a preliminary and correlational design. Though the measure is short, requiring only 2 to 3 minutes to complete, each subscale correlates strongly with an established and longer measure of the same or similar construct. Furthermore, five out of seven of these concurrent validity coefficients are as high or somewhat higher than those reported by Locke et al. (2011) for the subscales of the CCAPS-62. The only exceptions to this were the Eating Concerns ($r = .648$ for the CCAPS-62 vs. $r = .520$ for the CCAPS-34) and Alcohol Use ($r = .811$ for the CCAPS-62 vs. $r = .777$ for the CCAPS-34) subscales. Thus, the item reduction process described in Study 1 seems to have successfully preserved the factor structure of the larger instrument (with the exception of the Family Distress subscale, which is not assessed in the CCAPS-34), and the stand-alone short-form subscales demonstrated acceptably high-score internal consistency and strong concurrent validity evidence.

Study 3

The CCAPS-34 was specifically designed for use as a measure of psychological symptoms in a particular clinical setting (counseling centers) with a particular population (college students). In college counseling, one of the primary goals of an instrument is to assess change occurring over time due to treatment, rather than random fluctuations due to factors such as mood. The main purpose of Study 3 was to document the level of stability of CCAPS-34 subscale scores in a nonclinical undergraduate student population over 1-week and 2-week intervals, which are the most common intervals between counseling sessions in regular practice. Counseling is presumed to affect individuals' and group mean CCAPS-34 subscale scores but not every counseling client is expected to change at the same rate in counseling. This could introduce potential changes in group means, as well as affecting intraindividual stability in psychological symptoms (e.g., Fisher, Newman, & Molenaar, 2011). Therefore, counseling center clients would not be appropriate for assessing the stability of the CCAPS-34, and nonclinical college students are more appropriate, though the level of intra-individual stability in counseling would be of interest for future study.

Participants

Participants in Study 3 were 155 undergraduate students from the same university as in Study 2, and these participants completed the study for course credit. Of these, 101 participants were recruited to the 1-week group and 54 were recruited to the 2-week group. Only participants who did not report current psychological or psychiatric treatment and completed both assessments within their group were included in the analyses, leaving 86 participants in the 1-week group and 47 participants in the 2-week group.

The 1-week group's mean age was 18.78 years and 67 (66%) were first-year students, 18 (18%) were sophomores, 12 (12%) were juniors, and 4 (4%) were seniors. The sample was predominantly female (74%) and European American (78.2%). Only one participant reported currently receiving counseling and only four reported current psychiatric medication and so were removed from the sample. The mean age of the 2-week group was 18.50

Table 5. Test–Retest Stability of the CCAPS-34 Subscale Scores Over 1-Week and 2-Week Intervals

| | 1-Week Group (*n* = 86) | | | | | 2-Week Group (*n* = 47) | | | | |
| | Time 1 | | Time 2 | | | Time 1 | | Time 2 | | |
CCAPS-34 Subscale	*M*	*SD*	*M*	*SD*	*r*	*M*	*SD*	*M*	*SD*	*r*
Depression	0.705	0.796	0.650	0.795	.866	0.691	0.751	0.712	0.823	.864
Generalized Anxiety	1.048	0.857	0.961	0.809	.857	1.068	0.853	0.965	0.805	.850
Social Anxiety	1.596	0.877	1.573	0.855	.851	1.448	0.846	1.394	0.864	.805
Academic Distress	1.072	0.829	1.108	0.884	.794	1.245	0.864	1.218	0.790	.742
Eating Concerns	1.228	1.076	1.207	1.130	.815	1.296	1.093	1.290	1.070	.771
Hostility	0.586	0.627	0.641	0.700	.813	0.608	0.606	0.632	0.694	.751
Alcohol Use	0.918	0.816	0.911	0.919	.792	1.032	0.939	0.964	0.954	.781

Note: CCAPS = Counseling Center Assessment of Psychological Symptoms. All test–retest correlations were significant. *p*s < .05.

years, and 40 (74%) were first-year students, 8 were sophomores (15%), 4 were juniors (7%), and 2 were seniors (4%). This sample was also predominantly female (75.9%) and European American (90.7%). Three participants reported current counseling and four reported current psychiatric medication and so were removed prior to analysis.

Procedure

All study measures were completed online. After completing the informed consent, all participants in both groups completed a brief demographics questionnaire followed by the CCAPS-34. One week later, participants in the 1-week group were prompted by e-mail to complete the CCAPS-34 again. One week after this, participants in the 2-week group were prompted by e-mail to complete the CCAPS-34.

Results

Test–retest coefficients (Pearson product–moment correlations) are presented in Table 5. All coefficients were significantly different from 0 (*p* < .01). The coefficients for the 1-week test–retest group ranged from *r* = .792 (Alcohol Use) to *r* = .866 (Depression). For the 2-week group, test–retest coefficients

ranged from *r* = .742 (Academic Distress) to *r* = .864 (Depression).

Discussion

Overall, this study showed acceptable test–retest stability from the short-form scores of the CCAPS-34 in a nonclinical sample of college students who are not in treatment. Relatively high test–retest stability is expected in this population over these intervals, though mean levels of symptoms of psychological distress may change over the course of time. The relatively high levels of test–retest stability observed across most of the subscales offer initial evidence of stability in the CCAPS-34 subscales among students who are not in treatment. It may not be surprising that the two subscales with lower test–retest scores in both 1-week and 2-week intervals include Academic Distress and Alcohol Use. Individual scores on these subscales may be more strongly influenced by time-sensitive events in the lives of college students and may be more likely to fluctuate for some students. Overall, this study supports the use of the CCAPS in college students by providing evidence that nonclinical college students' scores on the CCAPS-34 are reasonably stable over 1-week and 2-week intervals without intervention. In counseling, change on the CCAPS-34

subscale scores may be indicative of meaningful shifts in psychological symptoms.

General Discussion

The development of the CCAPS instruments represents a decade of sustained and collaborative effort between researchers and college counseling centers to address the assessment needs of counseling centers. The goal of this collaboration is to create clinically relevant, empirically rigorous, multidimensional, standardized instruments of mental health symptoms and distress that are designed for college students and appropriate both for assessment and for outcome/monitoring purposes. Although the CCAPS-62 is being used at hundreds of institutions, feedback from the field indicated that its length was prohibitive for the purpose of treatment monitoring/outcome evaluation and that some centers, pressed for time, space, and resources, needed a briefer instrument for initial assessment.

This feedback dovetailed with a growing trend in the field of clinical assessment to minimize the length of assessment instruments and thus the burden on clients, clinicians, and researchers. Creating psychometrically sound short forms (that are representative of an original, longer measure in both content and psychometric properties) may have numerous clinical impacts, such as promoting expanded assessment of treatment services, increasing the volume and quality of information gathered from clients to inform treatment, and facilitating repeated measurement of mental health symptoms during treatment. This may also lead to improvements in clinical tools available to counselors, treatment effectiveness, and the understanding of both presenting concerns and the change process in college counseling.

The three studies described herein provide an overview of the development and initial validation of a short form of the CCAPS-62. Both CTT and IRT approaches were used to identify potential item sets for seven of eight subscales in the short form. Whereas a CTT approach is easy to implement and does not have stringent requirements for sample size, an IRT approach is preferable due to its ability to provide information estimates at each level of the latent trait. For the CCAPS-34, item selections were performed separately for each of the seven subscales—Depression, Generalized Anxiety, Social Anxiety, Academic Distress, Eating Concerns, Hostility, and Alcohol Use—to meet the unidimensionality assumption of the employed IRT model. To ensure that the resulting instrument was both empirically rigorous and clinically relevant, the content of each item was carefully evaluated by clinical experts in concert with monitoring of both item and subscale psychometric properties. Resulting short forms for each subscale were further examined by checking their projected internal consistency and marginal reliability estimates to ensure that they did not discard excessive amounts of information from the original subscale and retained reasonably high reliability for their scores. The results of the CCT and IRT item selections were further cross-validated using an independent sample of responses. The final short form CCAPS-34 is almost half the length of the CCAPS-62, has good discrimination power, and performs well within the college student population. Furthermore, when the CCAPS-34 was administered as a stand-alone instrument, its seven-factor structure was supported in a CFA, and scores from all seven subscales demonstrated evidence of concurrent validity with established measures of similar constructs within college students, more than acceptable internal consistency reliabilities, and acceptable test–retest stability over both 1- and 2-week intervals in a non-clinical college sample.

Limitations and Future Directions

A number of limitations are apparent that dovetail with the future development and validation of the CCAPS-34. First, although the CCAPS-34 is a multidimensional measure

that assesses a wide range of college student issues, it does not (and was not intended to) address all potential areas of concern. Instrument development, and especially short-form development, must balance the demands of comprehensiveness and brevity within the context of intended use. Because the CCAPS-34 was intended to be a short, multidimensional instrument, it should be used as but one source of clinical information and clinicians should strive to be aware of what is not assessed. This is also underscored by the fact that in this study, we have only investigated self-report measures, rather than incorporating observer or behavioral ratings. Although doing so is consistent with the primary means of collecting information in counseling centers, correlations among self-report data are potentially confounded by self-schema. Future research is necessary to determine the extent of this potential problem with regard to the CCAPS-34.

Although we consider the CCAPS-34 suitable for use in clinical and research settings based on the findings presented herein, additional studies are needed to further explore its psychometric validity in the college student population. For example, Study 2 provided strong initial support for concurrent validity of the CCAPS-34 subscales within a nonclinical college student population. Further research will be needed to replicate and extend evidence of concurrent validity in the counseling center clients and to examine the clinical validity of the CCAPS-34 as a measure of treatment efficacy while other studies will be needed to examine the instrument's performance within diverse groups. As another example, Study 3 provides strong initial evidence that the CCAPS-34 is stable over time in a nonclinical sample—a key step in developing an instrument that is sensitive to change. However, events associated with the ebb and flow of a typical college semester (arrival, start of classes, mid-term or final exams, spring-break, warm weather, etc.) may represent predictable variables that could affect the stability of CCAPS-34 scores in the college setting. Therefore, future research will be needed to carefully examine the performance of the instrument across the academic calendar in both clinical and nonclinical groups, for instance, examining factorial invariance across groups and measurement occasions.

Extensions of the CCAPS-34 are also possible. For instance, because of the intercorrelations between subscale scores, it is possible that higher order factor analysis may reveal the presence of hierarchical structure among scales. Finally, because the CCAPS-34 is a short multidimensional instrument, there may be a set of specific variables (e.g., readiness for change, functioning, coping/resilience, and working alliance) that could be used in conjunction with the CCAPS-34 to enhance the treatment process.

Conclusion

Given the promising initial evidence of reliability (internal consistency and test–retest) and validity (concurrent), the CCAPS-34 appears suitable for use in clinical and research settings with college students. The CCAPS-34 was conceptualized and designed specifically for the purposes of repeated measurement (e.g., monitoring treatment progress or outcome) and brief symptom assessment when time, space, or resources prohibit the use of the full version. The adequate reliability and validity evidence that is reported here, along with the minimal time required to complete it (2–3 minutes), provides strong initial support for these uses of the CCAPS-34.

Declaration of Conflicting Interests

Funding

References

Agliata, A. K., & Renk, K. (2009). College students' affective distress: The role of expectation

discrepancies and communication. *Journal of Child and Family Studies, 18*, 396–411. doi:10.1007/s10826-008-9244-8

Anton, W. D., & Reed, J. R. (1991). *College Adjustment Scales: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Benton, S. A., Robertson, J. M., Tseng, W., Newton, F. B., & Benton, S. L. (2003). Changes in counseling center client problems across 13 years. *Professional Psychology: Research and Practice, 34*, 66–72.

Baker, R. W., & Siryk, B. (1986). Exploratory intervention with a scale measuring adjustment to college. *Journal of Counseling Psychology, 33,* 31–38.

Beck, A. T., Epstein, N., Brown, G., & Steer R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology, 56*, 893–897.

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8*, 77–100.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 53–63.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.). *Testing structural equation models* (pp. 445–455). Newbury Park, CA: Sage.

Carr, T., Moss, T., & Harry, D. (2005). The DAS24: A short form of the Derriford Appearance Scale DAS59 to measure individual responses to living with problems of appearance. *British Journal of Health Psychology, 10*, 285–298.

Center for Collegiate Mental Health. (2010). *CCAPS 2010 user manual*. University Park: The Pennsylvania State University.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Coles, M. E., Gibb, B. E., & Heimberg, R. G. (2001). Psychometric evaluation of the Beck Depression Inventory in adults with social anxiety disorder. *Depression and Anxiety, 14*, 145–148.

Derogatis, L. R. (1994). *SCL-90-R: Administration scoring and procedures manual*. Minneapolis, MN: National Computer Systems.

Fisher, A. J., Newman, M. G., & Molenaar, P. C. M. (2011). A quantitative method for the analysis of nomothetic relationships between idiographic structures: Dynamic patterns create attractor states for sustained posttreatment change. *Journal of Consulting and Clinical Psychology, 79*, 552–563.

Fuqua, D. R., Leonard, E., Masters, M. A., Smith, R. J., Campbell, J. L., & Fischer, P. C. (1991). A structural analysis of the State-Trait Anger Expression Inventory. *Educational and Psychological Measurement, 51*, 439–446.

Garner, D. M., Olmsted, M. P., Polivy, J., & Garfinkel, P. E. (1982). The Eating Attitudes Test: Psychometric features and clinical correlates. *Psychological Medicine, 12*, 871–878.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347–360.

Hambleton, R. K. (1993). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). Phoenix, AZ: Oryx Press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Norwell, MA: Kluwer Academic.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177–189.

Hu, L.-T., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jöreskog, K., & Sörbom, D. (2006). LISREL 8.80 [Software program]. Chicago, IL: Scientific Software.

Kopta, S. M., & Lowry, J. L. (2002). Psychometric evaluation of the behavioral health questionnaire-20: A brief instrument for assessing global mental health and the three phases of psychotherapy outcome. *Psychotherapy Research, 12*, 413–426. doi:10.1093/ptr/12.4.413

Kraus, D. R., Seligman, D. A., & Jordan, J. R. (2005). Validation of a behavioral health treatment

outcome and assessment tool designed for naturalistic settings: The treatment outcome package. *Journal of Clinical Psychology, 61*, 285–314.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy, 3*, 249–258.

Lei, P., Wu, Q., DiPerna, J., & Morgan, P. (2009). Developing short forms of the EARLI numeracy measures: Comparison of item selection methods. *Educational and Psychological Measurement, 69*, 825–842.

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an Ant Colony Optimization Algorithm. *Multivariate Behavioral Research, 43*, 411–431.

Locke, B. D., Buzolitz, J. S., Lei, P-W., Boswell, J. F., McAleavey, A. A., Sevig, T. D., . . . & Hayes, J. A. (2011). Development of the counseling center assessment of psychological symptoms-62 (CCAPS-62). *Journal of Counseling Psychology, 58*, 97–109. doi: 10.1037/a0021282

Millon, T., Strack, S., Millon-Niedbala, C. M., & Grossman, S. D. (2008). Using the Millon College Counseling Inventory to assess student mental health needs. *Journal of College Counseling, 11*, 159–172.

Mintz, L. B., & O'Halloran, M. S. (2000). The Eating Attitudes Test: Validation with *DSM-IV* eating disorder criteria. *Journal of Personality Assessment, 74*, 489–503.

Newman, M. G., Kachin, K. E., Zuellig, A. R., Constantino, M. C., & Cashman-McGrath, L. (2003). The Social Phobia Diagnostic Questionnaire: Preliminary validation of a new self-report diagnostic measure of social phobia. *Psychological Medicine, 33*, 623–635.

O'Malley, K., Wheeler, I., Murphey, J., O'Connell, J., & Waldo, M. (1990). Changes in levels of psychopathology being treated at college and university counseling centers. *Journal of College Student Development, 31*, 464–465.

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe Crowne Social Desirability Scale. *Journal of Clinical Psychology, 38*, 119–125.

Robertson, J. M., Benton, S. L., Newton, F. B., Downey, R. G., Marsh, P. A., Benton, S. A., . . . Shin, K.-H. (2006). K-State Problem Identification Rating Scales for college students. *Measurement and Evaluation in Counseling and Development, 39*, 141–160.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100–114.

Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption–II. *Addiction, 88*, 791–804.

Spielberger, C. D. (1999). *STAXI-2: The State-Trait Anger Expression Inventory professional manual.* Odessa, FL: Psychological Assessment Resources.

Steer, R. A., Ranieri, W. F., Beck, A. T., & Clark, D. A. (1993). Further evidence for the validity of the Beck Anxiety Inventory with psychiatric outpatients. *Journal of Anxiety Disorders, 7*, 195–205. doi:10.1016/0887-6185(93)90002-3

Swindle, R., Cameron, A., & Rosen, R. A. (2006). 15-Item short form of the Psychological and Interpersonal Relationship Scales. *International Journal of Impotence Research, 18*, 82–88.

Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item analysis and test scoring using item response theory.* Chicago, IL: Scientific Software.

Bios

Benjamin D. Locke received a PhD from Boston College. He is currently the Associate Director of Clinical Services for Counseling and Psychological Services at Penn State University and the Executive Director of the Center for Collegiate Mental Health.

Andrew A. McAleavey is a doctoral candidate in Clinical Psychology at Penn State University. His research interests include quantitative methods of psychotherapy process and outcome.

Yu Zhao is a doctoral candidate in Educational Psychology at Penn State University. Her research interests include educational and psychological measurement and testing.

Pui-Wa Lei received a PhD from the University of Iowa. She is currently an Associate Professor at Penn State University. Her research interests include methodological issues of multivariate statistical techniques and applications of item response theory.

Jeffrey A. Hayes received a PhD from the University of Maryland. He is currently a Professor of Counseling Psychology at Penn State University. His research interests include therapist effects in psychotherapy, therapists' experience, and issues of diversity in the psychotherapy process.

Louis G. Castonguay received a PhD from Stony Brook University. He is currently a Professor of Psychology at Penn State University. His research interests include psychotherapy process, outcome, integration, and training.

Hongli Li received a PhD from Penn State University. She is currently an assistant professor of Research, Measurement, and Statistics in the department of Educational Policy Studies at Georgia State University. Her research interests include applied measurement in education and quantitative methods.

Robin Tate is a doctoral candidate in Educational Psychology at Penn State University.

Yu-Chu Lin is a doctoral candidate in Educational Psychology at Penn State University.