The Impact of Screen Format and Repeated Assessment on Responses to a Measure of Depressive Symptomology Completed Twice in a Short Timeframe

by

Patricia Wallis

B.A., University of Victoria, 2005

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
Department of Psychology
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Fall 2013

Approval

Name: **Patricia Wallis**

Degree: **Master of Arts (Psychology)**

Title of Thesis: ***The Impact of Screen Format and Repeated Assessment on Responses to a Measure of Depressive Symptomology Completed Twice in a Short Timeframe***

Examining Committee: **Chair:** Robert McMahon
Professor

Rachel Tanya Foualdi
Senior Supervisor
Associate Professor

Firstname Surname
Supervisor
Assistant/Associate/Professor

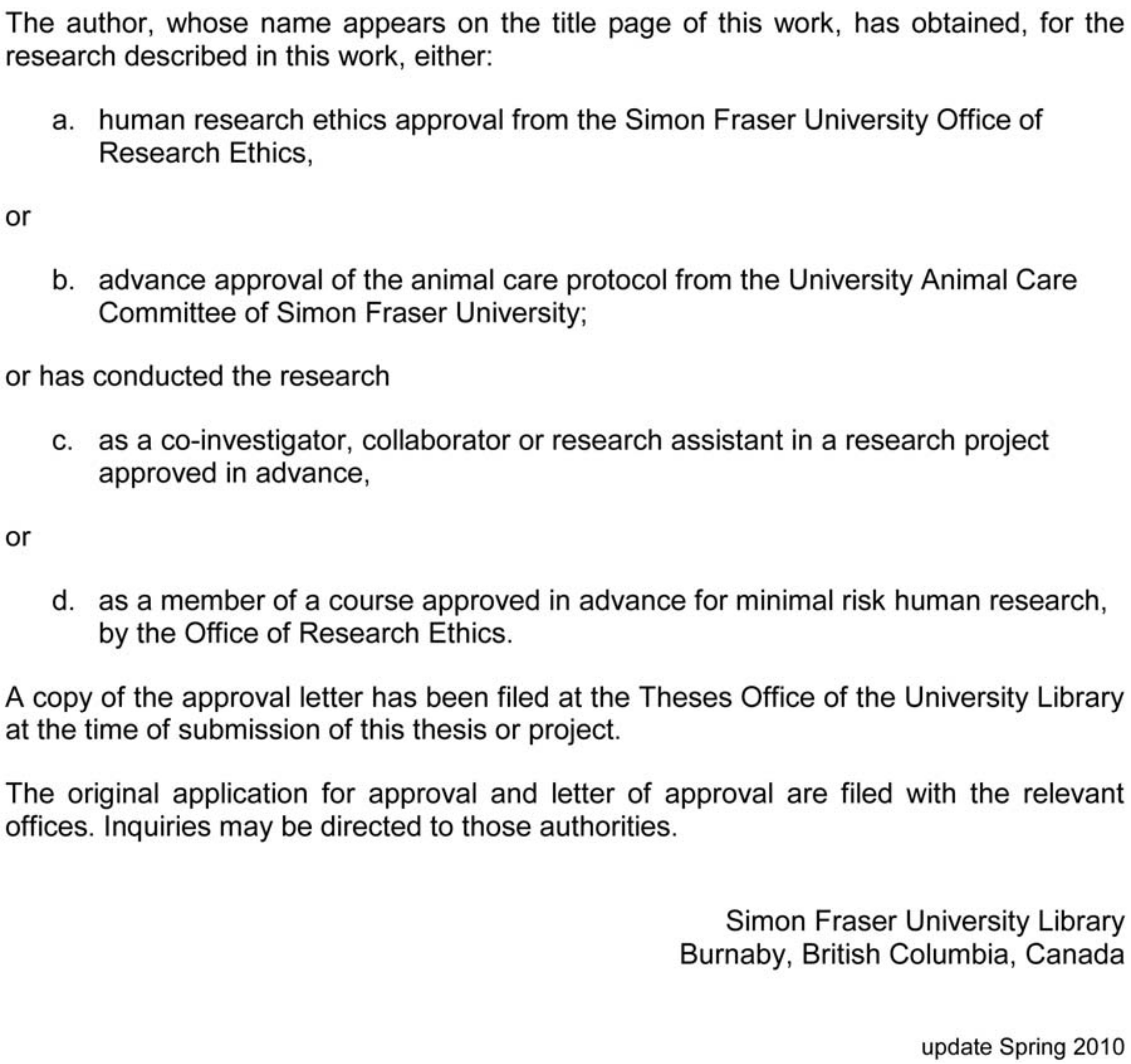
Allen Thornton
Supervisor
Associate Professor

Jeremy Biesanz
External Examiner
Associate Professor
Psychology
University of British Columbia

Date Defended/Approved: December 9, 2013

Partial Copyright Licence

Ethics Statement



The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics,

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University;

or has conducted the research

c. as a co-investigator, collaborator or research assistant in a research project approved in advance,

or

d. as a member of a course approved in advance for minimal risk human research, by the Office of Research Ethics.

A copy of the approval letter has been filed at the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

update Spring 2010

Abstract

The quality of psychology research produced and the policy developed based on this research are directly related to the accuracy of measurement. By conducting research that identifies the causes of error, it is possible to more accurately predict or minimize this error (Groves & Lyberg, 2010). In the present study, a repeated measures design was used to study the effect of screen format and repeated assessment on participant responses to a twenty item measure of depressive symptomology over the past week. There was no effect of format at the scale score and categorization level, however an effect of format was present for some subscales and items, but not others. Consistent with previous research (e.g., Arrindell, 2001), an effect of repeated assessment was present with participants reporting lower levels of depressive symptomology on the second assessment compared to the first assessment when considering overall composite scores. In addition this retest effect was present for categorizations based on composite scores, subscale scores, and almost half of the twenty items. The effect of screen format and repeated assessment on responses to the measure of depressive symptomology was relatively consistent for males and females and people with different self-reported levels of English fluency.

Keywords: Measurement; electronic questionnaires; retest effect; format effect; depressive symptomology; sex differences; language effects; item characteristics

Acknowledgements

This research was funded through the Social Sciences and Humanities Research Council Standard Research Grant to my supervisor, Rachel Fouladi. Without the council's support, this basic research into the occurrence of response distortions and factors influencing those distortions on a commonly used measure of depressive symptomology would not have been possible. I would like to thank my supervisor, Rachel Fouladi, for the mentorship she provided as well as the study participants, the research assistants, and my supervisory committee. In particular, I would like to acknowledge that this degree would not have been possible without the assistance of Anita Turner, the graduate program assistant. I would also like to thank my family and friends for their assistance and support during this process. Particularly my good friend Shawn Salsiccioli – I would like to thank him for being awesome.

Table of Contents

List of Tables

List of Figures

1. Introduction

Measurement is the “first building block of science” (Babbie, 1990, p. 20) and a foundation of psychological research and practice. The quality of psychology research produced and the policy developed based on this research are directly related to the accuracy of measurement. One perspective regarding the role of measurement is “to reach truth is the aim of knowledge and measurement is the operative means to get true data” (Mari, 2005 p. 260). However, measurement in psychology, like many other disciplines, is imperfect.

Self-report questionnaires are a frequently used method of measurement and assessment in both research and clinical practice. Electronic and web surveys have become more common as they can be less expensive compared to other modes, such as telephone or face-to-face surveys, can be a faster mode of data collection (Shannon & Bradshaw, 2002), and can lead to faster data processing (Kiesler & Sproull, 1986). Electronic data collection can have advantages in clinical research settings. For example, electronic data collection for practice-based research networks, where data are collected in clinical environments from multiple institutions, has the potential for improved data transfer to the centralized data facilities (Pace & Staton, 2005). Thriemer et al. (2012) in a fever surveillance study in Tanzania found that while the start-up costs associated with electronic data collection were higher, the cost of conducting the study was 25% less. Scharer et al. (2002) describe potential benefits of electronic data collection using personal digital assistants (PDA) to gather ongoing data from people with bipolar disorder for clinical use as the cost would be less compared to creating similar paper-and-pencil tracking journals.

Previous research has considered differences between paper-and-pencil and electronic modes of data collection (e.g., Fouladi, McCarthy, & Moller, 2002; McCabe, Boyd, Couper, Crawford, & Darcy, 2002). More recently, researchers have started to evaluate the way different features of electronic questionnaires (e.g., pictures on the screen, scrolling between items or not) impact the way participants respond to electronic questionnaires (e.g., Toepoel & Couper, 2011; Peytchev, Couper, McCabe, & Crawford, 2006; Mahon-Haft & Dillman, 2010). By conducting research which identifies the causes of error, it is possible to more accurately predict or minimize this error (Groves & Lyberg, 2010).

Measurement error is the difference between the value provided by a participant on a given variable and the true value of that variable, which is unknown. The observed value on a variable can be thought of as the combination of a participant's true score on that variable and measurement error. Measurement error includes both random and systematic error. Random error varies in an unpredictable way over repeated measurements (Joint Committee for Guides in Metrology [JCGM], 2012). In large samples, random error will tend to balance out in terms of estimates of the mean of a given variable, however will impact the variability (Niemi, 1993). Systematic error remains constant or varies in a predictable way over repeated measurements (JCGM, 2012), leading to measurement bias which can influence a variable mean (Niemi, 1993). The present study will consider two potential sources of systematic error in measurement using electronic self-report questionnaires: repeated assessment and questionnaire format.

In order for a respondent to answer a question optimally, a respondent must interpret the question, identify relevant information from their memory, use this information to form a single judgement, and translate that judgement into a response or response option (Krosnick, 1999). Each of these steps is complex and there is opportunity for error at each stage (Schwarz & Oyserman, 2001). When completing a survey interview (1995) or a questionnaire (1998), Schwarz suggests that participants will use the principles of cooperative conversation when interpreting an item and deciding on an appropriate response. These principles of cooperative conversation (Grice, 1975) include the maxim of quantity (i.e., provide enough information, but not more than is necessary), maxim of quality (i.e., information is believed to be accurate), maxim of relation (i.e., be relevant), maxim of manner (i.e., be clear). This means that participants may use information such as the numbers associated with response options or surrounding question items to provide information about question meaning or anticipated responses, thereby potentially influencing responding.

Following are sections providing an overview of questionnaire formatting, retest effects, the relationship between language fluency and gender on depressive symptomology, scoring and item characteristics. Each section contains background information and places the factor in the context of the current study. Following these sections is further description of the present study, including specific research questions.

1.1 Format

Computers are an increasingly frequent mode of data collection. One advantage of electronic questionnaires is the variety of screen formats available, some of which were not practical with paper-and-pencil questionnaires (e.g., a single item per screen). Understanding whether, and how, the format of a questionnaire affects responses is key in developing consistent measures that can be easily interpreted across studies. Similarly if electronic data collection is used in clinical settings for ongoing monitoring of client status/symptom level, the impact of screen design on client responses is important to understand in interpreting the data collected.

Some recommendations for web survey design have been made. For example, Reips (2002) comments on the importance of not setting up the survey where the neutral response is pre-selected because non-responses and neutral responses will be indistinguishable. Heerwegh and Loosveldt (2002) suggest that radio buttons are easier to use than drop-down boxes as participants required more time to complete the questionnaire with drop-down boxes and the drop-down box format had a significantly higher dropout rate for the study.

Differences in participants' responses have been found due to different formats. For example, Christian and Dillman (2004) presented response options horizontally across the screen with "excellent", "good", "poor" on the first row and "very good" and "fair" on the second row. Participants were more likely to select "very good" and less likely to select "good" for this format than when response options were listed vertically in a single list. Smyth, Dillman, Christian, and Stern (2006) presented participants with a question regarding financial support for school and provided a list of responses in a check-all-that-apply format. When responses from the list were categorized into subgroups with headings, participants were more likely to select at least one response from each sub-group than when the response options were listed as a single list. Hartley and Betts (2009) found when response options presented the positive response option label with the highest corresponding number on the left side of the screen, respondents' ratings were higher compared to other presented versions. Teopoel, Das, and van Soest (2008) also found numbers assigned to response option categories had an effect on responses. When response options ranged from 2 to -2, responses were different than when response options had other labels, as participants tended to not assign negative scores.

Christian and Dillman (2004) suggests that participants use the layout of a questionnaire as a source of information when selecting a response. This idea is supported by Stern's findings (2006) that when a check-all-that-apply list is separated into subsections participants are more likely to select at least one response from each subsection. Tourangeau, Couper, and Conrad (2004) suggest that respondents will use visual cues when interpreting a questionnaire, including item proximity. A group of items that are presented on a single screen may be interpreted as more related or asking about the same overall topic compared to individual items each presented on individual screens. Results are mixed whether participants interpret items as more related when presented together.

Tourangeau, Couper, and Conrad (2004) presented participants with eight items related to diet with a seven-point response scale ranging from "Agree" to "Disagree." Tourangeau et al. found when the items were presented on a single screen Cronbach's alpha was higher than when the items were presented on separate screens, indicating the correlation among items was higher when participants completed items on a single screen than when the items were presented on separate screens. Couper, Traugott, and Lamais (2001) presented participants with eleven items on attitudes on affirmative action, rated on a 5-point Likert scale, on either a single screen or with a single item on each screen. Couper et al. found no statistical differences between Cronbach's alpha depending on screen format, although the trend was in the expected direction, with a higher Cronbach's alpha among items presented together on a single screen compared to Cronbach's alpha among items presented one item per screen. Neither Tourangeau et al. nor Couper et al. commented on whether there were differences on reported attitudes to affirmative action or diet depending on survey format. Thorndike et al. (2009) presented the Beck Depression Inventory (BDI), Beck Anxiety Inventory, Quality of Life Index, and Montgomery-Asberg Depression Rating Scale twice to participants, once with a single item on the screen and once with all of the items on a single screen. Thorndike et al. found the factor structure and factor means were consistent for each of the measures between formats. Toepoel, Das, and van Soest (2009) found a higher number of items presented on the screen was associated with an increased likelihood of a participant skipping at least one item and also a greater number of skipped items compared to when fewer items were presented on the screen at once. Additionally, Toepoel et al. found having multiple items on the screen decreased the time taken to complete the survey, however participants' preference ratings of screen layout were lower for participants presented multiple items per screen. Similarly, after completing measures in both formats, Thorndike et al. found participants reported preferring a single item per screen compared to multiple items per screen.

1.1.1 Present study

In the present study, participants completed a self-report measure in two different formats. The first format is a single item per screen. The second format has multiple items on a screen with response options presented beside the items. These formats were selected because each has a different strength in terms of ease of use and potential differences in the way participants complete the measures from an information processing perspective. The one-item on a page format has little visual clutter. The multiple items and response options beside format is similar to the layout in traditional paper-and-pencil questionnaires. In addition, while previous research has evaluated the correlation between items in these two formats, few studies appear to have examined whether presentation format has an impact on reported scores or the factor structure of a measure. Studies on the effect of questionnaire format have been primarily conducted with between groups designs using randomization (e.g., Couper et al., 2001; Tourangeau et al., 2004; Christian & Dillman, 2004). In these designs, when differences are found between groups, the differences are attributed to the factor of interest (e.g., screen format). However, when differences are present there is no way to determine if the groups tested had identical theoretical score distributions on the construct measured in the first place. In the present study, each participant completed both formats. In this case, because the same people are completing each of the formats, there is only one group, thus there is only one theoretical score distribution on the construct measured. Any observed differences are due to factors other than pre-existing differences between groups.

1.2 Repeated Assessment

In some research and clinical settings, people complete the same self-report measure more than once. Examples include assessment of participants' change over the course of a given treatment in therapy outcome studies. When scores on self-report measures completed more than once are compared to evaluate change on a given construct, it is assumed that different reported scores are reflective of different levels of the construct. One would expect that people will report similar scores across multiple iterations when there is no theoretical reason for change in the level of the construct. For example, across a large number people, one would expect general psychiatric functioning to be the same at multiple assessments provided there is no external reason to expect a change (e.g., therapy, major event in society).

Previous research has shown that on some measures of negative mood participants report increased levels of functioning or decreased negative symtomatology when multiple assessments are conducted when there is no clear external reason for the reported change (e.g., Ormel, Koeter, & van den Brink, 1989; Sharpe & Gilbert, 1998; Arrindell, 2001). This retest effect of reported increased functioning across multiple assessments occurs primarily between Time 1 and Time 2 (Arrindell, 2001). In other words, participants reported level of functioning increases between Time 1 and Time 2, but at Time 2, Time 3 and subsequent assessments, participants generally report similar levels of functioning.

Researchers have shown this retest effect, where participants report increased functioning at Time 2 with no external treatment, in a number of measures. These include the Beck Depression Inventory (BDI) and Hamilton Depression Rating Scale (HAM-D) depression measures (e.g., Deardoff & Funabiki, 1985; Ahava et al., 1998), the

General Health Questionnaire (GHQ), a measure of general psychiatric functioning (Ormel et al., 1989), the SCL-90-R, which includes a number of subscales such as anxiety, agoraphobia, general psychological distress (Arrindell, 2001). Interestingly, this effect has not been found in measures of positive states (Arrindell, 2001; Sharpe & Gilbert, 1998).

This retest effect has been demonstrated in various populations and across varying time periods between assessments. Deardoff and Funabiki (1985), Ahava et al. (1998) demonstrated reported increased functioning across multiple assessments in undergraduate college students. Ormel et al. (1989) and Arrindell (2001) reported this retest effect in clinical populations, either inpatient or outpatient. In the study by Arrindell, time between assessments ranged from 11 to 350 days for inpatients and 3 months for outpatients. The time between assessments did not have an effect of the observed retest effect. Hatzenbuehler et al. (1983) observed the retest effect in college students where there were only hours between assessments. A study by Longwell and Truax (2005) had different findings than other studies reviewed here. Longwell and Truax assigned participants to complete the BDI either weekly, monthly or bimonthly for 9 weeks. Only participants who completed the questionnaire weekly showed the retest effect of reported increased functioning across the 9 week period. Additionally, the authors concluded there was an effect of frequency of assessment because there were differences in reported scores between the conditions at week 5 and week 9, where the participants in the different conditions had completed the assessment different numbers of times. This is different from other findings that suggest the primary change in reported functioning occurs between Time 1 and Time 2. Similar to previous work, Longwell and Truax concluded no effect of time as the reported scores were the same for the three conditions at the second assessment. This means reported scores were the same at Time 2 for the people who completed the assessment 1 week later, 1 month later, and 2 months later.

Researchers have proposed a number of explanations for this retest effect, however, there has been little research evaluating and comparing these different explanations. In general the explanations for the retest effect fall into one of two perspectives. The first perspective takes the position that there is a real change in functioning or symptomatology occurring between Time 1 and Time 2 and this change in functioning or symptomatology is accurately captured by the change in reported functioning in the measure. Essentially, there is a real change occurring and this is being reflected in the change in scores on the measures. The second perspective takes the position that the reported change in scores between Time 1 and Time 2 is a measurement error or artefact and does not represent a real change in functioning. Essentially, something about completing a self-report measure more than once leads people to report higher functioning even though there actual level of functioning has not changed.

1.2.1 Present study

Compared to previous work, this study adds a number of new elements. This retest effect has been observed on paper-and-pencil measures of negative mood. As the reason for this effect is not well understood, it is unclear whether this effect is present on electronic measures in addition to paper-and-pencil measures. The present study tests whether electronic questionnaires are also impacted by this retest effect.

An additional difference from prior studies is the presentation of the Center for Epidemiologic Studies – Depression scale (CES-D) twice in immediate succession. While previous work has demonstrated the retest effect over a number of time frames, it has not been evaluated when the measure was completed in immediate succession. While Swartz et al. (2007) presented the CES-D in different modes twice consecutively, they did not directly address the effect of repeated assessment on CES-D scores. However, an interaction between order of presentation and mode was present. Presenting the CES-D in immediate succession provides information regarding some of the explanations that have been put forward explaining the retest effect. An important feature of the CES-D is the type of items included. The CES-D asks about the frequency of specific behaviours or feelings over the past week. Essentially, if there are differences between scores taken within the same testing session, at least one assessment must be inaccurate because there has not been an opportunity for behaviours over the past week to change. If a retest effect is observed, it would provide evidence against explanations of this effect that suggest a real change on the construct measured is occurring.

1.3 Individual Characteristics

A variety of individual characteristics have been explored with regard to their association with response patterns on questionnaires. Examples of individual characteristics include demographic as well as personality characteristics. In the following, sex and English language fluency are considered with regard to their association with responses to questionnaires regarding depressive symptomology levels.

1.3.1 Sex

Females consistently report higher levels of depressive symptomology (Boticello, 2009; Hankin & Abramson, 2001; Wade, Cairney, & Pevalin, 2002; Kessler, McGonagle, Swartz, Blazer, & Nelson, 1993; Culbertson, 1997; Akhtar-Danesh & Landeen, 2007). Higher reported levels of depression in females have been found in adolescents (Hankin & Abramson, 2001; Wade, Cairney, & Pevalin, 2002) and this pattern continues into adulthood (Kessler et al., 1993). In addition, this phenomenon has been reported across cultures (Kuehner, 2003; Weissman et al., 1996; Maier et al., 1999). In the present study, the effect of gender is examined and controlled for because previous research has demonstrated different reported levels of depressive symptomology in males and females.

In addition, in a meta-analysis Voyer, Voyer, and Bryden (1995) found males and females have demonstrated different spatial abilities. This included spatial tasks that required the ability to determine spatial relations with distracting information present. Because this type of difference in spatial ability could affect the way a participant responds to a given questionnaire format, in this study the interaction between gender and the other factors are tested.

Gender is also included in item level analyses because of potential differences in item functioning depending on gender. Lange, Thalbourne, Houran, and Lester (2002) found women were more likely report somatic complaints (decreased food intake, hypersomnia, and low sex drive) and were more likely to worry about being poor compared to men reporting the same level of depressive symptoms on the Thalbourne's Manic-Depressiveness Scale. Differential item functioning was reported for the "crying" item on the CES-D by Gelin and Zumbo (2003) in a sample of 600 adults from northern British Columbia and by Cole, Kawachi, Maller, and Berkamn (2000) in a sample of 2340 community-dwelling seniors based on gender. Females had a higher level of endorsement for the "crying" item than males with the same reported level of depressive symptomology. Consistent with Gelin and Zumbo and Cole et al., Stommel et al. (1993) found higher levels of endorsement in women than men for the "crying" item with comparable levels of depressive symptomology. Stommel et al. also reported the "talked less" item was had lower levels of endorsement in females than males with the same reported level of depressive symtomology in a sample of 1212 adults.

1.3.2 Language Fluency

Increased English language fluency is associated with lower levels of depressive symptomatology as measured on the CES-D (Rumbaut, 1994). A longitudinal study by Beiser and Hou (2001) of Southeast Asian refugees in Canada found that after ten years in Canada, English fluency was a predictor of depression. Lee and Chen (2000) found that competence speaking English was associated with lower self-reported depressive symptomology in immigrant Chinese adolescents living in Canada. In 2007, Dao, Lee, and Chang reported that among 112 Taiwanese international students in the United States, students with lower perceived English fluency reported higher levels of depressed feelings. In a sample of 83 Vietnamese immigrant and refugee women, women with greater English fluency had lower levels of depressive symptoms (Brown, Schale, & Nilsson, 2010).

In addition to reported level of depression, response style is also related to language fluency and whether a measure is completed in a participants' native language. A response style is a response bias that is consistent for an individual, reflecting an individual style of responding (Jackson & Messick, 1958). A study by Harzing (2006) considered the relationship between response style and whether the questionnaire was completed in the participant's native language. Business students (*N*=1,581) from 26 countries completed questionnaires with 5-point response scales in

English and the native language of the country where the data was collected. Harzing found that students with higher levels of English language fluency had higher levels of extreme responding and decreased levels of mid-point responding. In addition, participants completing the questionnaire in their native language were more likely to engage in extreme responding, while participants completing the questionnaire in English were more likely to select middle responses. Participants with higher levels of English fluency completing the measure in English responded in way that was similar to participants completing the measure in their native language (i.e., more extreme responses). These are consistent with findings by Gibbons, Zellner, and Rudek (1999) that selecting extreme responses to a 5-point likert scale was more common when participants responded to items in their native language compared to a second language.

Cultural differences have also been found in response styles. For example, Hammura, Heine, and Paulhus (2008) found differences in response styles between 158 Canadian university students of East-Asian heritage compared to Canadian university students of European heritage. Participants were categorized based on whether they spoke a European or East-Asian language at home. Participants of East-Asian heritage demonstrated more ambivalent and moderate response styles when completing the Big Five Inventory and fifteen additional items both with a 7-point response format. Additionally, in a study of 95 American and Korean college students, Lock and Baik (2009) found that Korean college students demonstrated a more acquiescent response style compared to American college students.

Using data from three large (at least 1700 participants) market research studies in six European Union countries, Greece, Italy, Spain, France, Germany, and the United Kingdom, van Herk, Poortinga, and Verhallen (2004) compared differences in acquiescent responding and extreme responding between participants from different countries. Van Herk, Poortinga, and Verhallen concluded higher levels of acquiescent and extreme responding were found in respondents from Greece compared to respondents from the other EU countries. Additionally, higher levels of acquiescent and extreme responding were found in respondents from Spain and Italy compared to respondents from the United Kingdom, Germany, and France.

Presently there is not a consensus in the explanation of differing response styles between cultures. Hammura, Heine, and Paulhus (2008) findings support the theory that cultural differences in dialectical thinking, an openness to holding apparently contrary beliefs or ideas, may explain response style differences between Canadian students of East-Asian heritage compared to European heritage. Smith (2004) proposed that differences in acquiescence bias across nations is relatively stable and has substantive cultural meaning. Using data from published studies that included samples from at least 34 countries, Smith found higher levels of acquiesence bias in countries with higher levels of family collectivism for personally relevant items and lower levels of uncertainty avoidance for items related to a respondents' perceptions of society. These findings highlight not only the importance of participant characteristics, but also item characteristics; differences in response patterns depending on the item content is an example of item features differentially impacting the way that participants respond.

Hui and Triandis (1989) found that extreme responding, the frequency of selecting the endpoints of a scale, was more frequent in Hispanics compared to non-Hispanics when responses options were presented as a 5-point scale. However, there was no difference in the level of extreme responding between Hispanics and non-Hispanics when response options were presented using a 10-point scale. While people may differ in their inherent tendency to follow a particular response style and response style appears to vary across cultures, it is possible that response style may be encouraged or discouraged by situational factors, such as questionnaire format (e.g., Hui & Triandis, 1989).

In cognitive assessments, the Cultural-Language Interpretive Matrix is a framework where the impact of two factors, linguistic demand and cultural loading, of an assessment are considered in selecting and interpreting assessment tools (Flanagan & Ortiz, 2001). It is possible that linguistic demand and cultural loading impact responses to measures in addition to cognitive assessments. For example, on a multi-item measure some items may be more difficult to read and understand which could impact the way in which people with different levels of English fluency respond.

In the present study, English language fluency is examined and controlled for when assessing reported depressive symptomatology because, as demonstrated in previous literature, it is expected that participants reporting lower levels of language fluency will report higher levels of depressive symptomatology. In addition, the interaction effect of language fluency with the effect of format and repeated assessment of depressive symptomology is considered because previous literature on response styles, suggests the way people tend to respond to questionnaires may differ between cultures. Similar to Hammuara et al. (2008) language fluency would be used as a proxy for acculturation. Characteristics of items, such as reading difficulty (Appendix A), are also considered in the present study.

1.4 Scoring and Item Characteristics

Different approaches to scoring measures, including the CES-D, are used in research. Typically composite scores on the CES-D are computed by summing across the 20 items. However, other scoring procedures are found in the literature. For example, computing subscale scores (e.g., Nikolova, 2012) or categorizing participants based on composite scores (e.g., French, 2012; Patten, Lavorato, & Metz, 2005). In the present study, each of these three approaches to scoring (composite scores, subscale scores, categorization) are considered when assessing the relationship between CES-D scores and repeated assessment, format, language fluency and sex. Effects of repeated assessment or format may differentially impact items. In the present study, item level analyses are conducted to explore whether differences that may be occurring due to format and multiple assessments are driven by certain items, while responses to other items are not affected by format and multiple assessment.

Item characteristics, such as topic and readability, have also been associated with responding. For example, respondents tend to underreport behaviour on items regarding sensitive or socially undesirable topics (e.g., illicit drug use) and tend to over-report on items regarding socially desirable behaviour (e.g., voting) (Tourangeau & Yan, 2007). In a study of 115 items Velez and Ashworth (2007) found items with higher reading grade levels were associated with higher rates of midpoint responding. Additionally, decreased item clarity, as rated by seven raters, was associated with higher rates of midpoint responding. Specifically, in addition to item readability, the effects of several coder-identified item characteristics on response consistency across consecutive assessments are examined in the present study.

1.5 Overview of Present Study and Research Questions

The current study is an examination of questionnaire format and repeated assessment which are two factors that may impact the accuracy of measurement using self-report questionnaires. Two formats of the target measure are presented to each participant. English language fluency and sex are also included as factors because of their potential relationship to levels of reported depressive symptomology.

Depressive symptomology is the construct that is the focus of the present study because it is a frequently measured construct in psychological research, often assessed on multiple occasions (e.g., Vahdaninia, Omidvari, & Montazeri, 2010; Watkins, Baeyens, & Read, 2009; Kroenke et al., 2011). As the retest effect of higher levels of functioning reported at assessment on the second assessment has been demonstrated in this construct, depressive symptomology is a reasonable choice for further examination of this effect.

The Center for Epidemiological Studies-Depression (CES-D) is the specific measure examined in the present study. The CES-D is a 20-item measure of depressive symptoms over the past week. Participants are instructed to rate each item based on how many times they felt a given way or engaged in a given behaviour during the past week. An example CES-D item is “I was bothered by things that don’t usually bother me.” The measure uses a 4-point ordered response scale with response options ranging from “rarely (less than 1 day)” to “most of the time (5 to 7 days)”. Each item is scored from 0 to 3. Four items are reverse coded. Higher summated composite scores (theoretical range: 0-60) indicate higher levels of depressive feelings. Following Rushton, Forcier, and Schectman (2002), participants can be categorized into minimal (0-15), mild (16-23), or moderate/severe levels of depressive symptomology (≥24). Traditionally CES-D scores over sixteen are considered suggestive of significant depressive symptomology (Weissman, Sholomskas, Pottenger, Prusoff, & Locke, 1977), however Roberts, Lewinsohn, and Seeley (1991) suggest that a cutoff of 24 may be more likely to detect Diagnostic and Statistical Manual of Mental Disorders (DSM) defined depression in adolescents.

Radloff (1977) identified a four factor structure in the CES-D. The four factors are Depressed Affect, Positive Affect, Somatic Symptoms, and Interpersonal Problems. In a meta-analysis of 28 studies including either an exploratory factor analysis or principal components analysis of the CES-D, Shafer (2006, p. 134) concluded "the results were clear and highly consistent with the initial factor analyses conducted by Radloff." The CES-D has demonstrated good reliability and validity in adolescents and adults (Radloff, 1991).

Two forms of the CES-D were presented to participants. These formats include one item per screen and multiple items per screen with response options beside the items. Participants completed the two forms of the CES-D twice in immediate succession. Differences between a participant's score on the first and second version of the CES-D are considered an indication of measurement error because the CES-D asks about frequency of specific behaviours and feelings over the past week. As the two versions of the CES-D are presented within a single testing session, there is little opportunity for additional real instances of these behaviours and feelings to occur.

In terms of repeated assessment, the goal of the present study is to determine whether the retest effect occurs in a single testing session with this particular measure and to provide information for or against some explanatory theories of the retest effect. In terms of format, the goal of the present study is to test whether there are differences between participants' responses to these two formats and to provide information on the way layout and visual cues impact the way participants respond to questionnaires. The use of a repeated measures design to address issues of questionnaire formatting complements and adds to previous studies, which used a between subjects design. The research questions for the current study are detailed below.

1. i. Do participants respond differently to the same measure of depressive symptomology completed in two different formats on the computer?

   ii. Is the effect of format on CES-D responses different for people with different individual characteristics (i.e., level of English fluency and gender)?

2. i. Do participants respond differently to the same measure of depressive symptomology when the measure is completed twice consecutively?

   ii. Is the effect of repeated assessment on CES-D responses different for people with different individual characteristics (i.e., level of English fluency and gender)?

3. Do findings for Research Question 1 and Research Question 2 vary as a function of the way depressive symptomology is considered (i.e., assessing composite scores, assessing CES-D categorization, at the item level)?

4. Are features of the item (e.g., readability) predictive of participants' changes in item responses?

2. Method

2.1 Participants

This study was conducted in compliance with university ethics guidelines and with Human Subjects Approval from the institutional review board. Participants included 954 undergraduate university students recruited through the Department of Psychology Research Participation System (RPS) at Simon Fraser University and advertisements in *The Peak*, a Simon Fraser University student newspaper. Participants recruited through the RPS received credit towards their undergraduate psychology course. The only inclusion criterion was a willingness to participate in a one-hour session completing a number of questionnaires. Each participant provided informed consent before completing the study.

Data from three participants were dropped due to high levels of missing data, leading to a sample size of 951. Initially, participants with missing data on either measured independent variable, sex and English fluency, were included in analyses. However, in the generalized estimating equations analyses, no solutions were found when these participants were included. When these eleven participants were dropped from the analysis, solutions were found. In order to keep the sample consistent across all analyses, the results of reported analyses do not include the eleven participants who did not provide information on either the sex or language fluency item. The sample size included in subsequent analyses is 940.

Three hundred twenty-four participants were male (M_{age}*=19.95,* SD_{age}=2.94), 616 participants were female (M_{age}*=19.46,* SD_{age}=2.31). Participants primarily self-identified as Asian (55%) or Caucasian (29%). The remaining sixteen percent self-identified as another ethnicity including first nations, biracial, and other. Half of participants (50%) identified English as their first language. On an English fluency item, 50.5% of participants self-identified as “very fluent, English is my first language”, 23.7% self-identified as “more fluent than my first language”, 14.3% self-identified as the “same fluency as my first language”, and 11.5% self-identified as “less fluent in English than my first language”.

In order to determine an appropriate sample size, a power analysis was conducted using PASS software, Power Analysis and Sample Size (Hintze, 2008). Four factors and all interaction terms were included in the model. The variables included two within factors with two levels each and two between groups factors, one with two levels and one with four levels. Results indicate a total sample size of 928 is necessary for a power level of at least .85 for each effect to detect small effects (d=.2 for all interaction terms and d=.1 for main effects). These effect sizes are consistent with previous research on the primary variables of interest in the proposed study, retest effects (e.g., d_z=.2 for both depression subscales included in Arrindell, 2001; η^2=.1 for depression scales included in Sharpe and Gilbert, 1998) and format effects (e.g., $.1 \leq \eta^2 \leq .2$ for different response option presentations in Hartley and Betts, 2009).

2.2 Measures

2.2.1 Demographics Instrument

Demographic information was collected using a multi-item instrument including questions about sex, age, ethnic/racial identification, and English language fluency. The English language fluency item was “How fluent are you in English?” Response options were a 4-point ordered response scale ranging from “Very fluent, English is my first language” to “Less fluent than my first language.”

2.2.2 Center for Epidemiological Studies-Depression (CES-D)

The CES-D (Radloff, 1977) is a 20-item measure of depressive symptoms over the past week. Table 1 provides a list of CES-D items. Table 2 presents the CES-D items grouped by subscale. Figures 1 and 2 display the two formats of the CES-D presented. In one format, a single item was presented on each screen. In the second format, multiple items were presented on the screen with response options presented beside the items.

Overall composite scores are created by summing across the twenty CES-D items. The theoretical range of the total scale score is 0-60. Subscale scores are created by summing across the items included in the subscale. The four items on the positive subscale are reverse coded prior to computing the subscale score; thus, similar to overall composite scores and other three subscales, higher scores on the positive subscale indicate higher levels of depressive symptomology.

Overall composite CES-D scores were computed using a pro-rated composite score for participants who completed 80% of the items. A pro-rated subscale score was computed for a participant on a given subscale if the participant completed five of seven items for the somatic subscale and the depressed affect subscale, three of four items on the positive affect subscale, and one of two items on the interpersonal subscale.

The total score on the CES-D has demonstrated good internal consistency, reliability and validity for use in adolescents and adults (Radloff, 1991). Cronbach's alpha for the measure in the present study was .88 (95% CI: .87-.89) on the first assessment and .89 (95% CI: .88-.90) on the second. Cronbach's alpha for each subscale on the first and second assessment were: somatic symptoms .63, .69; depressed affect .84, .87; positive affect .81, .85; and interpersonal problems .58, .71. Tables 3, 4, and 5 include 95% confidence intervals for Cronbach's alpha for the overall composite and subscale scores for each timepoint and for each format.

2.3 Procedure

Participants completed a battery of electronic questionnaires in a lab at Simon Fraser University. Each participant completed the measures on a laptop computer with a mouse. After participants provided consent, research assistants read participants the instructions. Instructions indicated participants would complete a series of electronic questionnaires, including some questionnaires in different formats. First, participants completed a series of demographic items. Next, participants completed the CES-D twice in immediate succession. The CES-D was formatted differently for the presentations, one form with a single item on each screen and the other form with multiple items on the screen and response options presented beside the items.

2.4 Design

Each participant completed two formats of the CES-D successively to evaluate the effect of format and repeated assessment on participants' responses on the CES-D. The order of the formats was counterbalanced. This study has four factors, two within (format and time of assessment) and two between (English fluency and gender) subjects factors; first, time of assessment, which has two levels (Time 1 and Time 2); second, format, which has two levels (one-item per screen and multiple items per screen with response options beside items); third, English fluency, with four levels (very fluent, English is my first language; more fluent than my first language; same fluency as my first language; and less fluent than my first language); fourth, gender, with two levels (male and female). Format and time of assessment were crossed. English fluency and gender were measured variables. The dependent variable was CES-D responses. CES-D responses were considered in the following ways; first, as summated composite scores; second, by category for level of depressive symptomology -- following Rushton, Forcier, and Schectman (2002) participants were categorized into minimal (0-15), mild (16-23), or moderate/severe levels of depressive symptoms (≥24) based on summated composite CES-D score. Finally, subscale and item level analyses were conducted.

3. Analysis

A general overview of the analytic approach followed by detailed analytic strategies for each of the research questions is presented here.

3.1 Descriptives and Diagnostics

Demographic characteristics of the sample were described. Standard descriptive statistics were computed for measured variables used in subsequent analyses. Diagnostics for assumption checking were conducted prior to further analysis. For example, for the linear mixed models, the normality assumptions were assessed using q-q plots of residuals. Additionally, scatterplots were created to evaluate the relationship between CES-D and other continuous variables to evaluate whether the relationship is linear.

3.2 Means

General linear mixed model (LMM) analyses, using restricted maximum likelihood estimation, were conducted to examine mean levels of depressive symptomatology as a function of time of assessment, format, English fluency, and gender. The repeated measures dependency structure was selected from appropriate models based on fit using the Bayesian Information Criteria (BIC). Main effects and interaction effects were examined as appropriate using *F*-tests of parameters. Follow-up tests controlling for set-wise type I error were conducted as necessary. All necessary follow-up pairwise comparisons in the study were conducted using a Bonferroni correction in SPSS; the *p*-values reported for these follow-up pairwise comparisons have been multiplied by a Bonferroni multiplier to adjust for multiple tests and are denoted as

p_{mc}. The same sets of analyses were conducted with each of the four CES-D subscale scores and item level responses as the dependent variables. Ninety percent confidence intervals of the non-centrality parameters (NCP) are provided as an indicator of effect size (Steiger & Fouladi, 1997) for terms in the LMM models.

3.3 Categorizations

Generalized estimating equations (GEE) analyses were conducted to test whether the proportion of people categorized as having different levels of depressive symptoms (minimal, mild, moderate/severe) varied as a function of time of assessment, format, English fluency, and gender. The working correlation matrix was selected based on fit using the quasi-likelihood under the independence model criterion (QIC), a modification of the Akaike information criteria for GEE models. Main effects and interaction effects were examined as appropriate using Wald χ^2 tests of parameters. Follow-up tests controlling for set-wise type 1 error were conducted as necessary. Ninety percent confidence intervals of the non-centrality parameters are provided as an indicator of effect size for terms in the GEE models.

In terms of controlling for type I error, the tests on composite scores, categorizations, subscale scores, and item responses were considered as four families of tests; family-wise type I within each of these sets of tests was controlled as appropriate. A Bonferroni correction was used for each of the first three families of tests and an alpha of .01 was used for each analysis in the family of tests on item responses. For this reason, alpha was set to .05 for the composite score model, .025 for the two GEE models of categorizations, .0125 for the four subscale models, and .01 for the item response models.

LMM and GEE were selected to flexibly model the within subject correlation due to repeated measurement on the CES-D; additionally, missing data on a given variable does not result in the deletion of cases. LMM and GEE analyses were conducted using SPSS 17.0. Confidence intervals of noncentrality parameter (NCP) estimates were obtained using the Noncentral Distribution Calculator (NDC) (Steiger, n.d.); Appendix B gives corresponding η^2, η, f^2, and f values.

3.4 Test Structure

Test structure was considered in two ways. First, test structure was considered by evaluating whether the expected four factor structure of responses to the CES-D (Depressed Affect, Positive Affect, Somatic Symptoms, and Interpersonal Problems) fit the data at each timepoint overall[1] and within each group (e.g., males and females) at each timepoint. This was done with confirmatory factor analysis (CFA) using MLR estimation, a maximum likelihood estimator that is robust to multivariate non-normality of observations and can be used with missing data, in Mplus 7.11. Goodness-of-fit was assessed using corresponding Root Mean Square Error of Approximation (RMSEA) values and confidence intervals (Steiger & Fouladi, 1997), Standardized Root Mean Square Residual (SRMR) values, and χ^2 tests. While a variety of cutoff scores have been suggested to indicate fit, Hu and Bentler (1999) suggest when RMSEA and SRMR are used together, an RMSEA <.06 and an SRMR <.09 generally indicate good fit. Browne and Cudeck (1993) suggested an RMSEA value of <.08 would indicate a reasonable fit. Standardized residuals and model parameter estimates were evaluated using standard normal z-values.

Second, test structure was considered by evaluating whether the measurement model linking CES-D indicator items to the four factors (Depressed Affect, Positive Affect, Somatic Symptoms, and Interpersonal Problems) is identical between group conditions. This was done by testing for measurement invariance using multigroup factor analysis (MGFA). Following Muthén and Muthén (2009), three models were run for each multigroup analysis to determine the level of measurement invariance; (1) a configural model where the structure of the model is specified, but the factor loadings and intercepts can vary between groups; (2) a model specifying weak factorial

[1] A single factor model was also run for each timepoint, however the four factor model had better fit and was the only model considered in subsequent analysis. The four factor model was run in two ways for responses at both timepoints. 1) The four factor model was considered with the superordinate factor of depressive symptomology included. 2) The four factor model was considered where the superordinate factor was not included in the model. The model fit for these two models was the same at both timepoints. In subsequent analysis, the four factor model was run without the higher order factor of depressive symptomology included.

invariance where factor loadings are constrained as equal between groups, but intercepts are not; and (3) strong factorial invariance where both intercepts and factor loadings are constrained as equal between groups. For each multigroup analysis, differences in chi-square values between the more restrictive model and less restrictive model were tested to determine whether fixing the model to specify measurement invariance led to a more poorly fitting model. The difference in chi-square values was tested using a chi-square difference test for testing nested models using scaled chi-square values because MLR estimation was used (Satorra & Bentler, 1999, as described on the Mplus website, n.d.). If measurement invariance was not supported between groups, modification indices were evaluated. The comparisons included differences in fit between the one-item per screen and multiple items per screen formats on the each assessment (first and second), differences in fit between males and females on each assessment, and differences in fit between participants who indicated they were very fluent in English and participants who indicated they were less than very fluent in English on each assessment.

A similar approach, following Muthén and Muthén (2010), was also used for the test of measurement invariance between the first and second assessments, where increasingly restrictive measurement models were tested. In the first model, measurement invariance was not specified; in the second, invariance of factor loadings for each item on the first assessment is fixed as equal to the factor loading for the corresponding item on the second assessment. In the final model, both factor loadings and intercepts are constrained as equal for corresponding items on the first and second assessments. However, because of the repeated nature of the data, the model used for testing measurement invariance was the model used for testing for measurement invariance in growth curve models. Figure 3 depicts the structure of the model. Forty indicators were used; twenty from the first assessment and twenty for the second. Eight latent variables were specified representing the four subscales at each timepoint. Each of the latent variables was allowed to correlate with each of the seven other variables. Difference tests of scaled chi-square values would have been used to test for differences between the nested models, however due to model fit issues, these tests were not possible. An alternative model, discussed in the results section, was used and difference tests of scaled-chi square values were conducted on this revised model.

3.5 Item Characteristics

First, items were coded for a number of features including readability, referring to others, reverse coded or not, and item subscale membership. Readability was evaluated using Flesch-Kincaid grade level, which is computed based on the average word and sentence length. Flesch-Kincaid grade level values were computed using Word, 2003. Appendix A presents the formula for computing Flesch-Kincaid grade level. Other characteristics, such as whether the items refer to the participant or other people as well (e.g., I felt depressed versus I felt that I was just as good as other people), were coded by two separate raters.

Next, GEE was used to test whether the features of the item identified predict the likelihood participants' responses to the same item differed on the two testing occasions. The dependent variable was categorical: did a participant's response to the same item differ or was it the same on both occasions. Participants' responses to an item that were identical on both occasions were coded as zero, while responses that were not identical were coded as one. GEE analysis was selected to flexibly model the correlated nature of the data because each participant responded to multiple CES-D items and to allow for the binary dependent variable. The link function was binary logistic and the working correlation matrix was selected based on QIC fit values. Main effects and interaction effects were examined as appropriate using Wald χ^2 tests of parameters. Follow-up tests controlling for set-wise type I error were conducted as necessary.

3.6 Research Questions and Brief Summary of Corresponding Analyses

1. i. Do participants respond differently to the same measure of depressive symptomology completed in two different formats on the computer?

   To test whether there are differences in reported mean CES-D scores for the two formats the main effect of format using LMM was tested. This analysis was conducted separately with composite scores, subscale scores, and item responses as the dependent variable. GEE analyses were used to test whether the proportion of participants categorized as having different levels of depressive symptoms varies as a function of format. CFA was used to determine whether the structure of the responses to the CES-D for each format presented fit with the expected four factor structure and whether the factor structures are different between formats by testing for differences in the parameter values.

ii. Is the effect of format on CES-D responses different for people with different individual characteristics (i.e., level of English fluency or gender)?

To test whether the relationship between format and CES-D score is different for males and females or people with different levels of English fluency, the interactions between format and gender, between format and English fluency, between format and time, as well as the three and four-way interactions were included and tested as necessary in both the LMM and GEE models.

2. i. Do participants respond differently to the same measure of depressive symptomology when the measure is completed twice consecutively?

To test whether there are differences in reported mean CES-D scores for the two time points, the main effect of time using LMM was tested. This analysis was conducted separately with composite scores, subscale scores, and item responses as the dependent variable. GEE analyses were used to test whether the proportion of participants categorized as having different levels of depressive symptoms varies as a function of completing the CES-D twice consecutively. CFA was used to determine whether the structure of the responses to the CES-D at each time point fit with the expected four factor structure and whether the factor structures are different between assessments by testing for differences in the parameter values.

ii. Is the effect of repeated assessment on CES-D responses different for people with different individual characteristics (i.e., level of English fluency and gender)?

To test whether the relationship between time and CES-D score is different for males and females or people with different levels of English fluency, the interactions between time and gender, between time and English fluency, between time and format, as well as the interaction between time, gender and English fluency were included and tested as necessary in both the LMM and GEE models.

3. Do findings for Research Question 1 and Research Question 2 vary as a function of the way depressive symptomology is considered (i.e., assessing composite scores, assessing CES-D categorization, or at the item level)?

   A descriptive synthesis of the results of research questions one and two was conducted to evaluate whether the results of the research questions vary depending on the way depressive symptomology was considered (i.e., CES-D scores, subscales, categorization, items). A graphical overview of patterns of findings was presented as well as a tally of whether effects were significant or not or whether there were changes in effects or not depending on the way depressive symptomology was considered.

4. Are characteristics of the item (e.g., readability) predictive of participants' changes in item responses?

   LMM analysis was conducted to determine whether differences in CES-D scores depend on features of the items identified (e.g., readability). GEE analysis was conducted to test whether the features of the item identified (e.g., readability) predict the likelihood participants responses to the same item will differ on the two testing occasions.

4. Results

In the following section, descriptive statistics are presented. Next, model selection is described, followed by a description of tables summarizing findings from LMM and GEE models. Next, the main effects of sex and English fluency from LMMs, GEE models, and CFAs, which are not directly addressed by the four research questions, are presented. Then the specific results from each research question, based on LMM, GEE, and CFA, are presented.

4.1 Descriptive Statistics

Table 3 presents descriptive statistics (mean, median, min, max, standard deviation, skew, and kurtosis, as well as Cronbach's alpha) for CES-D composite scores at each timepoint and for each format. Tables 4 and 5 present descriptive statistics (mean, median, min, max, standard deviation, skew, kurtosis, and Cronbach's alpha) for each of the four subscales at each timepoint and for each format. Table 6 presents descriptive statistics for each of the twenty CES-D items at both timepoints. Table 7 presents descriptive statistics for each of the twenty CES-D items for each of the screen formats. Table 8 presents mean scale scores on the first and second assessment for each screen format. The Pearson correlation between composite scores on the first and second assessments was .963 (95% CI=.959-.967). The Pearson correlation between composite scores when the CES-D is presented with a single item per screen and presented with multiple items per screen was .957 (95% CI=.952-.962).

Table 9 presents descriptive statistics of scale scores for males and females for each format and for the first and second assessment. Table 10 presents the descriptive statistics of scale scores for participants indicating different levels of English fluency for each format and for both assessments. Tables 11 through 14 present descriptive statistics of subscale scores for males and females and for participants indicating different levels of English fluency. Tables 15 through 22 present descriptive statistics of item values.

4.2 LMM and GEE Model Selection for Research Questions 1-3

4.2.1 Linear Mixed Models

First, the repeated covariance matrix type was selected using BIC fit values for the full model including all interaction terms (time of assessment, format, sex, English language fluency, and all interaction terms) with the dependent variable of CES-D composite scores. The compound symmetric matrix had the best fit and was selected for further linear mixed model analysis.

Next, the terms to include in each linear mixed model were determined. The terms of interest in the research questions were the main effects of time of assessment, format, sex, English language fluency, and the interactions between time of assessment and sex, time of assessment and English language fluency, format and sex, and format and English language fluency. To determine which additional terms to include, linear mixed models were run for each dependent variable (CES-D composite scores, each CES-D subscale, and each CES-D item) using the full model including all interaction terms.

A series of LMMs were run for each dependent variable with higher order interactions dropped in stages. Interactions were dropped in this way because the number of people at each level of each measured variable (sex and English fluency) was not the same. First, for each dependent variable, the full model including all terms of interest in the study and all interaction was run. Next, the models were run without the 4-way interaction. Subsequently, the 3-way interactions were dropped, then 2-way interactions that were not of interest in the study; finally a model with only the main effects included was run. Comparing results across models for each dependent variable, no additional interaction terms were identified as statistically significant when interaction terms were dropped in blocks compared to when statistically significant interaction terms were identified from the full model.

Two sets of models were used in the final interpretations of results. First, models for each dependent variable including terms of interest and additional higher order terms if the term was statistically significant (p<.05) in the full model or if the term was a lower order interaction for a higher order interaction that was statistically significant. For example, for a particular dependent variable, if the interaction between sex, time of assessment, and format was statistically significant the 2-way interaction of time of assessment and format would be included in the analysis regardless of whether it was statistically significant in the full model. The 2-way interactions between sex and time of assessment and sex and format were already added to the model because they are terms of interest for specific research questions. Figure 4 displays which terms were included in the final model for each dependent variable. As noted previously, dropping interaction terms in blocks did not impact the terms identified for these models.

The second set of models was used to test the main effects of sex, English fluency, format, and repeated assessment. The models used for interpreting main effects included only the main effects and no interaction terms. Again, this was done because of the unbalanced nature of the measured variables.

4.2.2 Generalized Estimating Equations

First, the type of model for each of the GEE analyses was selected. For the 2-group categorization where scores that were less than 16 were categorized in the lower depressive symptomology group, a binary logistic model was used. For the 3-group categorization where participants with scale scores of less than 16 were in the lower depressive symptomology group and participants with scores greater than 23 were in the higher depressive symptomology group, an ordinal logistic model was used. Next, the working correlation matrix was selected based on results from the 2-group categorization model. QIC fit values were the same for three working correlation matrices, autoregressive one, exchangeable, and unstructured, so an autoregressive one working correlation matrix was selected.

Next, the terms included in each of the final models were selected using the same approach as for the LMMs. The terms of interest were included in the GEE analysis for the 2-group categorization and the 3-group categorization. To determine which additional terms to include, GEE analysis for the 2-group and 3-group categorizations were run including the full model with all interaction terms. In subsequent analyses, higher order terms were dropped out in blocks (i.e., 4-way, then 3-way, then 2-way interactions not of interest). No higher order interaction terms were statistically significant in any of these models. Interaction terms in addition to the terms of interest would have been included for a dependent variable if the term was statistically significant (p<.05) in the full model or if the term was a lower order interaction for a higher order interaction that was statistically significant. However, none of the additional interaction terms were statistically significant in either GEE model, so both final GEE models included only terms of interest.

Similar to the LMM analyses, two sets of models were used for interpretation. The first set of models for each categorization approach included the terms of interest. The second set of models included the main effects only.

4.3 Overall LMM and GEE Model Results for Research Questions 1-3

Tables 23, 24, and 25 present the results from the main effects LMMs. Tables 26, 27, and 28 present LMM results for the models including main effects and interactions. Table 29 presents the results from main effects GEE models. Table 30 presents the results from the GEE models including all terms of interest.

The next section describes the main effects of sex and English fluency from the LMM and GEE models as well as from MGFA models. The results from the main effects of sex and English fluency are presented separately because they were not specifically addressed by the research questions, but were included in the models because the interactions of sex and English fluency with format and repeated assessment were addressed in specific research questions. Subsequently, results from these models are presented in the section corresponding to the relevant research questions.

4.4 Sex and Language Fluency Main Effects in LMM, GEE, and CFA

4.4.1 Main effects of sex

Consistent with previous literature, mean CES-D scores for females were statistically significantly higher than the mean CES-D score for males, where higher scores indicate higher levels of reported depressive symptomology, $F_{(1,\ 935)}$=4.65, *p*=.031 (M_{sex} presented on Table 9). Consistent with composite score results, females also had a statistically increased likelihood of being in a higher depressive symptomology category in both the two group categorization, *Wald* $\chi^2(1)$=5.47, *p*=.019, and three group categorization, *Wald* $\chi^2(1)$=5.31, *p*=.021. Table 31 shows percent of males and females in each categorization group. However, statistically significant sex differences were only present in the LMM analyses on one of the four subscales ($M_{subscales}$ presented on Table 11). Depressed affect scores were higher for women compared to men, $F_{(1,\ 935)}$=22.41, *p*<.001. No differences were present between subscale scores for males and females on the somatic symptoms, positive affect, and interpersonal problems subscales, *ps*>.0125, as presented in Table 24. With the item level analyses, for items 1 (Som), 10 (Dep), 17 (Dep), and 18 (Dep), the mean item response for females was higher compared to the mean item response for males, *ps*<.0125, as presented in Table 24 (M_{items} presented on Tables 19 and 20). No other items yielded statistical differences in the item level analyses, *ps*>.0125. Item wording is presented in Tables 1 and 2.

In terms of test structure, Table 32 presents MLR chi-square tests of model fit as well as corresponding RMSEA and SRMR values for the four factor measurement model. Based on RMSEA values ($\text{RMSEA}_{T1\ Male}$=.053, 90% CI=.044-.062; $\text{RMSEA}_{T1\ Female}$=.052, 90% CI=.046-.058; $\text{RMSEA}_{T2\ Male}$=.062, 90% CI=.054-.070; $\text{RMSEA}_{T2\ Female}$=.062, 90% CI=.056-.068) and SRMR values ($\text{SRMR}_{T1\ Male}$=.051, $\text{SRMR}_{T1\ Female}$=.044, $\text{SRMR}_{T2\ Male}$=.059, $\text{SRMR}_{T2\ Female}$=.051) the four factor model appears appropriate. Table 32 presents the results of chi-square difference tests of difference in fit between the configural invariance model and increasingly constrained measurement models. The results indicate that on both the first ($\chi^2(16)_{\text{config vs weak}}$ =128.21, *p*<.001) and second assessment ($\chi^2(16)_{\text{config vs weak}}$ =118.65, *p*<.001), measurement was not invariant between males and females. Results support neither weak nor strong factorial invariance between males or females on the first or second assessment.

Based on previous research and evaluation of residuals and modification indices, measurement invariance was tested for the four factor CES-D model after items 10 (Dep) and 17 (Dep) were dropped. Results from these models are presented in Table 33. Results indicate measurement invariance was not present between males and females on the either the first or second assessment when items 10 (Dep) and 17 (Dep) are not included in the model.

4.4.2 Main effects of English fluency

Self-rated English language fluency was associated with composite CES-D scores, $F_{(3,\ 935)}$=9.70, *p*<.001. Table 34 presents results from follow-up pairwise comparisons. Pairwise comparisons indicate composite scores for participants who indicated they were very fluent in English were lower compared to participants who reported having the same fluency in English as their first language (EM_{Diff}=2.82, p_{mc}=.008) and participants who reported being less fluent in English than their first language (EM_{Diff}=4.64, p_{mc}<.001), where lower composite scores indicate lower levels of reported depressive symptomology. Reported mean difference scores are the differences in the estimated marginal means from the model including the main effects of sex, English fluency, format, and repeated assessment. In addition to reporting higher levels of depressive symptomology compared to participants who reported being very fluent in English, those who reported being less fluent in English than their first language also reported higher levels of depressive symptomology compared to participants who were more fluent in English than their first language (EM_{Diff}=3.12, p_{mc}=.015).

Consistent with composite score results, English fluency was also associated with the likelihood of being in a higher depressive symptomology category in both the two group categorization, *Wald* $\chi^2(3)$=30.44, *p*<.001, and three group categorization, *Wald* $\chi^2(3)=33.30$, $p<.001$. Table 35 shows percent of participants at each level of self-rated English fluency in each categorization group. For the two group categorization with the cutoff score of 16, a higher proportion of participants who identified as very fluent in English were in the lower depressive symptomology category compared to (1) participants who identified as more fluent in English than their first language (p_{mc}=.048), compared to (2) participants who identified as having the same fluency in English as their first language (p_{mc} =.005), and (3) compared to participants who identified as having less fluency in English compared to their first language (p_{mc}<*.001*). The proportion of participants in the lower depressive symptomology category was not different in any other pairwise comparisons (p_{mc} >.05)

As presented in Table 24, English fluency was associated with scores on the depressed affect subscale and positive affect (*ps*<.0125), but not with scores on the somatic symptoms and interpersonal problems subscales (*ps*>.0125) ($M_{subscales}$ presented on Tables 13 and 14). As presented in Table 25, for eleven of the twenty CES-D items (Som: 1, 7, 13; Pos: 4, 8, 12, 16; Dep: 6, 9, 14, 18), self-reported English fluency was associated with item response (M_{items} presented on Tables 15 and 16). Table 36 presents results of follow-up pairwise comparisons. For items 6 (Dep) and 14 (Dep), main effects of English fluency are present, however English fluency is included in statistically significant higher order interactions, so these effects are discussed in subsequent sections.

In testing for measurement invariance between participants with different levels of self-reported English fluency, the four possible English fluency ratings were grouped into two groups. Based on results from LMMs, the main differences in language fluency groups appeared between participants who were very fluent in English and other self-reported levels of English fluency and between participants who rated themselves as less fluent in English than their first language and other self-reported levels of English fluency. The number of participants in the less fluent in English than their first language was 108. Because of the relatively smaller group size for the less fluent in English group, this group was not tested as a separate group in the MGFA. Rather it was combined with the same fluency in English as first language, and more fluent in English than their first language so that measurement invariance was tested between participants who self-identified as very fluent in English and participants who did not self-identify as very fluent in English.

Table 37 presents the results presents MLR chi-square tests of model fit as well as corresponding RMSEA and SRMR values for the four factor measurement model. Based on RMSEA values ($RMSEA_{T1\ VF}$=.059, 90% CI=.053-.066; $RMSEA_{T1\ notVF}$=.062, 90% CI=.056-.069; $RMSEA_{T2\ VF}$=.066, 90% CI=.060-.073; $RMSEA_{T2\ not\ VF}$=.072, 90% CI=.065-.088) and SRMR values ($SRMR_{T1\ 1st\ VF}$=.052, $SRMR_{T1\ not\ VF}$=.051, $SRMR_{T2\ VF}$=.051, $SRMR_{T2\ not\ VF}$=.060) the four factor model appears to have a fair fit. For each of the four CFA models the SRMR value is <.09. For the two CFA models (one for participants who identified as very fluent in English and one for participants who did not identify themselves as very fluent in English) for responses on the first assessment, the RMSEA values are close to .06 and the 90% CI for RMSEA values is <.06. For the two CFA models for responses on the second assessment RMSEA values were between .06 and .07. While perhaps not indicating a good model fit, the RMSEA values are <.08, which Brown and Cudeck (1993) suggest indicates a fair fit.

Table 37 presents the results of Chi-square difference tests of difference in fit between the configural invariance model and increasingly constrained measurement models. The results indicate that on both the first ($\chi^2(16)_{weak}$ =32.68, *p*=.008) and second assessment ($\chi^2(16)_{weak}$ =41.13, *p*<.001), measurement was not invariant between participants who identified as very fluent in English and participants who did not identify as very fluent in English. Results support neither weak nor strong factorial invariance between the two English fluency groups on the first or second assessments.

Based on modification indices of the CFAs for each of the two language fluency categories, the tests for measurement invariance were run again after dropping items 10 (Dep) and 17 (Dep). Results from these models are presented in Table 38. Results indicate measurement invariance was not present between participants who self-identified as very fluent in English and participants who did not self-identify as very fluent in English on the either the first or second assessment when items 10 (Dep) and 17 (Dep) are not included in the model.

As one of the goals of the study was to evaluate the effect of different conditions on different scoring approaches to the CES-D, standard scoring was used in subsequent analyses despite issues of measurement invariance. Results from each research questions are presented in the following sections.

4.5 Research Question 1 – Format and Individual Characteristics

i. Do participants respond differently to the same measure of depressive symptomology completed in two different formats on the computer?

ii. Is the effect of format on CES-D responses different for people with different individual characteristics (i.e., level of English fluency or sex)

4.5.1 Composite Scores

Results from the linear mixed model indicate there were no differences in CES-D composite scores between the two screen formats, $F_{(1,\ 938)}$=.16, *p*=.689, NCP 90% CI=0-3.85. Additionally, there were no statistically significant interactions of format by sex or format by English fluency, indicating that the finding of no differences in scores depending on format was consistent across males and females and across participants with different levels of English fluency, $F_{S*F(1,\ 930)}$=1.83, *p*=.177; $F_{EF*F\ (3,\ 930)}$=.17, *p*=.916. Table 3 presents mean scale scores for each screen format.

4.5.2 Categorizations

Tables 29 and 30 present results from GEE models. For both the 2-group and 3-group categorizations, using GEE, there were no differences in the proportion of participants categorized at each level depending on the screen format, *Wald* $\chi^2(1)_{\text{2-group}}$=1.48, *p*=.223; *Wald* $\chi^2(1)_{\text{3-group}}$=1.38, *p*=.239. Table 39 shows percent of participants in each categorization group for each screen format. Additionally, there were no interactions between format and sex or format and language fluency for either categorization approach, *Wald* $\chi^2(1)_{\text{2-group } S*F}$=.04, *p*=.838; *Wald* $\chi^2(1)_{\text{3-group } S*F}$ =.19, *p*=.660; *Wald* $\chi^2(3)_{\text{2-group } EF*F}$=4.76, *p*=.191; *Wald* $\chi^2(3)_{\text{3-group } EF*F}$ =4.93, *p*=.177.

4.5.3 Subscale Scores

Results from LMMs regarding the effect of format on CES-D subscale scores were mixed. Two of the four subscales showed a difference in mean subscale scores depending on format, however the direction of the effect was different for the two subscales. Table 5 presents mean subscale scores for each screen format. For the somatic subscale, higher scores, indicating higher levels of reported depressive symptomology, were reported on the one-item per screen format compared to the multiple items per screen format, $F_{(1, 938)}$=9.74, *p*=.002. In contrast to the somatic subscale, higher scores on the positive subscale were reported for the multiple items per screen format compared to the one item per screen format, $F_{(1, 938)}$=28.74, *p*<.001. Items on the positive affect subscale are reverse coded before scoring; therefore consistent with the other three subscales, higher scores are associated with higher levels of reported depressive symptomology. For the depressed affect and interpersonal subscales no differences in subscale scores between the one-item per screen and multiple item per screen formats were present, $F_{\text{dep } (1, 938)}$=1.62, *p*=.204; $F_{\text{int } (1, 938)}$=1.35, *p*=.246. Confidence intervals of non-centrality parameters for results from each subscale are presented in Table 24.

The effect of format on subscale scores was not different for males and females or for participants with different self-rated English fluency. For each of the four subscales, there was no interaction of format by sex or format by English fluency, *ps*>.0125 for all interaction terms, as presented in Table 27.

4.5.4 Items

Results from LMM analyses testing the effect of format on participant responses to CES-D items indicated no differences in participants' responses for the majority of items. For eighteen items, no effect of format was found, indicating there were no differences in participants responses to an item depending on the format of the CES-D, *ps*>.01, as presented in Table 25. However, for items 12 (Pos) and 16 (Pos) mean participant item responses were higher, where higher scores indicate higher reported depressive symptomology, when the CES-D was presented in the multiple items per screen format ($F_{Item\ 12(1,\ 925)}$=8.02, *p*=.005 and $F_{Item\ 16\ (1,\ 930)}$=14.64, *p*=.001). Table 7 presents mean item responses for each screen format.

Generally, there were no differences in the effect of format on CES-D scores for males and females and participants with different levels of English language fluency, *ps*>.01, as presented in Table 28. However, for item 14 (Dep) a sex by English fluency by format interaction was present, $F_{(3,\ 916)}$=5.18, *p*=.001. When the relationship between English fluency and format on responses to item 14 (Dep) are evaluated separately for males and females, a main effect of format is present for males ($F_{(1,\ 318)}$=4.23, *p*=.040), but not females ($F_{(1,\ 603)}$=.01, *p*=.924). However, an effect of English fluency was present for females ($F_{(3,\ 610)}$=3.29, *p*=.020), but not for males ($F_{(3,\ 320)}$=1.81, *p*=.145). An interaction between English fluency and format was present for both males ($F_{(3,\ 318)}$=3.30, *p*=.021) and females ($F_{(3,\ 603)}$=2.77, *p*=.041). When evaluating the relationship between English fluency and responses to item 14 (Dep) for each format for males, differences in responses to item 14 (Dep) were present for the one item per screen format ($F_{(3,\ 319)}$=3.00, *p*=.031), but not for the multiple items per screen format ($F_{(3,\ 318)}$=.852, *p*=.466). Follow-up pairwise comparisons using a Bonferroni correction indicate for males on the one item per screen format, participants reporting less fluency in English than their first language had higher item responses to item 14 (Dep) compared to participants who were native English speakers (EM_{Diff}=.46, p_{mc}=.029). In contrast, when evaluating the relationship between English fluency and responses to item 14 (Dep) for each format for females, differences in responses to item 14 (Dep) were present for the multiple items per screen format ($F_{(3,\ 606)}$=4.84, *p*=.002), but not for the one item per screen format ($F_{(3,\ 610)}$=1.66, *p*=.175). Follow-up pairwise comparisons using a Bonferroni correction indicate for females on the multiple item per screen format, participants reporting less fluency in English than their first language had higher item responses to item 14 (Dep) compared to participants who were native English speakers (EM_{Diff}=.40, p_{mc}=.005) and participants who were more fluent in English than their first language (EM_{Diff}=.34, p_{mc}=.044).

In addition, while there was no main effect for either format or repeated assessment for item 15 (Int), a repeated assessment by format interaction was present, $F_{(1,\ 930)}$=7.04, p=.008. Follow-up tests indicate no differences in mean item response depending on screen format on either the first assessment ($F_{(1,\ 936)}$=3.32, p=.069), or the second assessment ($F_{(1,\ 935)}$=2.28, p=.131). The interaction was likely present because although not statistically different, on the first assessment, the response option beside format scores were higher compared to the one item per screen, but for the second presentation the responses to the one item per screen form were higher than scores on the multiple item per screen form.

4.5.5 Test Structure

First, the four factor measurement model was fit using CFA separately for the one item per screen format and the multiple items per screen format for responses on the first assessment and separately on the second assessment. Although chi-square tests of model fit do not indicate exact fit, the RMSEA and SRMR values indicate the model fit is adequate (all RMSEA values <.07 and all SRMR values <.06 as presented in Table 40).

RMSEA values for the MGFA indicate the four factor measurement model fit when configural invariance was specified for the one item per screen and multiple items per screen formats on both the first assessment (RMSEA=.062, 90% CI=.057-.066) and the second assessment (RMSEA=.068, 90% CI=.063-.072). Table 40 presents the Chi-square tests of model fit and RMSEA values for the MGFA when weak factorial invariance and strong factor invariance are specified on the first assessment and the second assessment. On the first assessment, results from testing increasingly specified measurement models indicate no decrement in fit was present ($\chi^2(16)_{weak}$ =6.39, p=.983; $\chi^2(20)_{strong}$ =23.97, p=.244) indicating measurement invariance between the two formats was present on the first assessment. On the second assessment, weak factorial invariance was present ($\chi^2(16)_{weak}$ =22.53, p=.127), however a decrease in model fit was present when constraints for strong factorial invariance were added ($\chi^2(20)_{strong}$ =40.54, p=.004). On the second assessment, weak factorial invariance was established for the one item per screen format and multiple items per screen format, however strong factorial invariance was not established.

Follow-up exploratory analyses of the strong factorial invariance model on the second assessment were conducted. When the intercepts of items 5 and 7 were not constrained as equal between the two formats, there was no statistical decrease in model fit between the weak factorial invariant model and the strong factorial invariant model, ($\chi^2(18)_{\text{strong}}$ =16.75, *p*=.540) model. Results from the follow-up model with the intercepts of items 5 and 7 not constrained are presented in Table 41. This suggests strong factorial invariance is present between the two formats except for items 5 and 7.

4.6 Research Question 2 – Repeated Assessment and Individual Characteristics

i. Do participants respond differently to the same measure of depressive symptomology when the measure is completed twice consecutively?

ii. Is the effect of repeated assessment on CES-D responses different for people with different individual characteristics (i.e., level of English fluency or gender)?

4.6.1 Composite Scores

LMM analysis indicate composite CES-D scores were higher on the first assessment compared to the second assessment, $F_{(1,\ 938)}$=157.66, *p*<.001, NCP 90% CI=117.52-203.50, where higher scores indicate higher reported levels of depressive symptomology. Table 3 presents mean scale scores on each assessment. There were no differences in this effect depending on sex or level of English language fluency, $F_{S*RA(1,\ 930)}$=.953, *p*=.329; $F_{EF*RA\ (3,\ 930)}$=2.17, *p*=.091.

4.6.2 Categorizations

For the 2-group categorization, repeated assessment had an impact on the categorization, Wald $\chi^2(1)=25.98$, $p<.001$. On the first assessment, a higher proportion of participants (.404) were categorized in the greater than or equal to 16 compared to the proportion categorized in the greater than or equal to 16 category on the second assessment (.355). Similarly for the 3-group categorization, repeated assessment had an impact on categorization, Wald $\chi^2(1)=18.78$, $p<.001$. Compared to the first assessment, at the second assessment a higher proportion of participants were in the less than 16 category and a lower proportion were in the between 16 and 23 and greater than 23 categories. For example, on the first assessment 60% of participants were categorized with a composite score of less than 16, while on the second assessment 65% of participants were categorized with a composite score of less than 16. Table 42 presents the number and percentage of participants in each category on the first assessment and the second assessment. The relationship between repeated assessment and categorization was consistent for males and females and participants with different reported levels of English fluency for both the two and 3-group categorizations, *ps*>.025 for all interaction terms, as presented in Table 30.

4.6.3 Subscale Scores

Consistent with results from composite CES-D scores, for each subscale participants reported lower CES-D scores on the second assessment compared to the first assessment, indicating lower reported depressive symptomology on the second assessment compared to the first. Table 4 presents the mean subscale score on each assessment. Participants reported higher levels of depressive symptomology on the first assessment compared to the second assessment on the somatic subscale ($F_{(1, 938)}=90.08$, $p<.001$), depressed affect subscale ($F_{(1, 938)}=80.51$, $p<.001$), positive affect subscale ($F_{(1, 938)}=7.42$, $p=.007$), and interpersonal subscale ($F_{(1, 938)}=15.63$, $p<.001$). Confidence intervals of non-centrality parameters are presented in Table 24.

The effect of repeated assessment on subscale score was not different for males and females or participants with different levels of self-rated English fluency. For each of the four subscales, there was no interaction between repeated assessment and sex and repeated assessment and English fluency, *ps*>.0125 for all interactions, as presented in Table 27.

4.6.4 Items

For ten of the twenty items (Som: 1, 2, 5, 7, 13; Dep: 3, 6, 10, 14; Int: 19) there was an effect of repeated assessment on mean item response, *ps*<.01, as presented in Table 25. For nine of these ten items (Som: 1, 2, 5, 7, 13; Dep: 3, 10, 14; Int: 19) this relationship was not different for males and females and participants with different levels of reported English fluency, *ps*>.01 for all interactions, as presented in Table 28. For each of these eight items with an effect of repeated assessments and no repeated assessment by sex or repeated assessment by English fluency interaction, the direction of the effect was consistent; participant scores on an item were higher on the first assessment compared to the second assessment, where higher scores are associated with higher levels of reported depressive symptomology. Table 6 presents mean item response for each assessment.

For item 6 (Dep), the relationship between repeated assessment and reported item score was different between participants with different reported levels of English language fluency, $F_{(3,\ 925)}$=5.01, *p*=.002. Table 43 presents the mean responses to item 6 (Dep) for participants at each level of self-rated English fluency. Reported scores on the second assessment were lower compared to the first assessment for participants who identified as (1) more fluent in English than their first language, $F_{(1,\ 221)}$=9.95, *p*=.002, participants who identified as (2) having the same fluency in English as my their language, $F_{(1,\ 131)}$=10.61, *p*=.001, and participants who identified as (3) having less fluency in English than their first language, $F_{(1,\ 107)}$=12.51, *p*=.001. For item 6 (Dep), no differences in item responses on the first assessment compared to the second assessment were present for participants who identified as native English speakers, $F_{(1,\ 471)}$=3.09, *p*=.080.

4.6.5 Test Structure

First, the fit of the four factor measurement model was established on the first assessment (χ^2(164)=689.25, *p*<.001; RMSEA=.058, 90% CI=.054-.063; SRMR=.046) and the second assessment (χ^2(164)=833.97, *p*<.001; RMSEA=.066, 90% CI=.062-.070; SRMR=.053). While chi-square values do not indicate exact model-data fit, SRMR and RMSEA values indicate the measurement model does generally fit as RMSEA values are below .08 and SRMR values are below .09.

Figure 3 presents the structure of the model used to test for measurement invariance. Results from the model were not appropriate for interpretation because the latent variable covariance matrix was not positive definite. This was likely because the scores on the first assessment were so highly correlated with scores on the second assessment ($r_{comp1,comp2}$=.963, $r_{som1,som2}$=.912, $r_{dep1,dep2}$=.945, $r_{pos1,pos2}$=.903, $r_{int1,int2}$=.833). The model-estimated correlations between each latent variable on the first assessment and the corresponding latent variable on the second assessment were greater than one for each of the four pairs of latent variables corresponding to the four subscales.

Various changes to the specified model were made in attempting to specify a model with a positive definite latent variable covariance matrix. First, the model was run with residuals between each item on the first assessment free to correlate with the corresponding item on the second assessment. Correlations greater than one were still present between latent variables in the estimated model. In the next model, correlations between the latent variables were fixed to .96. No solutions were found for the model. Models were run with different Mplus estimators including MLM and ML. Because no interpretable model was found for the model where measurement invariance was not specified, it was not possible to test for changes in model fit in the more constrained models specifying measurement invariance using the initially proposed model.

Two additional models were considered in an attempt to find a solution for a model to test measurement invariance over time. First, the superordinate latent factor of depressive symptomology was added for each timepoint in addition to the four latent subscale factors for each timepoint. The latent variable covariance matrix was not positive definite for this model. Next, a single factor measurement model at each timepoint was considered instead of the four factor model, where only a single latent factor of depressive symptomology for each timepoint was included in the model. The latent variable covariance matrix was not positive definite for this model.

Because none of the previous models produced usable solutions, a different approach was used. A four factor model was run with the forty items as indicators for four latent variables. Unlike the previous models considered, the latent variable representing the somatic symptoms subscale was considered the underlying construct for items on the somatic symptoms subscale on both the first and second assessment. The same approach was used for each of the other three latent variables representing the three other subscales. In addition, residuals of corresponding items were free to correlate. When residuals on corresponding time one and time two variables were freed to correlate, the model fit notably better (RMSEA=.046, 90% CI=.044-.047; SRMR=.052 compared to RMSEA=.118, 90% CI=.117-.120; SRMR=.085). Figure 5 presents the structure of this model. Each latent variable represents the aspect of depressive symptomology that items on the subscale are designed to measure on both the first and second assessment, as in theory the same construct is being measured at both timepoints.

As the model depicted in Figure 5 had a usable solution, increasingly constrained versions of this model were tested to evaluate measurement invariance across repeated assessments. Four increasingly constrained models were considered. In each model, variances for each of the four latent factors were set to one. Three of the models corresponded to constraints that were planned with the initially proposed models (e.g., factor loadings fixed as equal for corresponding items on the first and second assessment). An additional model was considered where all relevant factor loading pairs were constrained to be equal, but intercept pairs were constrained as equal only for items where no main effect of repeated assessment in LMMs of item responses was detected in research question 2 (items identified in Table 25 or Figure 7). Table 44 presents fit indices for each of these models as well as chi-square tests of difference in fit for increasingly constrained models.

The fit, using RMSEA values and 90% confidence intervals and SRMR values, is similar for the model where neither factor loadings nor intercepts are constrained, the model where factor loadings are constrained, and the model where factor loadings are constrained and intercepts are partially constrained. SRMR and RMSEA values are slightly different for the model where factor loadings are constrained and intercepts are fully constrained compared to the values for the other three models. However, the confidence intervals of RMSEA values for the model where factor loadings are constrained and intercepts are fully constrained overlap with the confidence intervals of RMSEA values for the other three models. This suggests the fit may be as good for the fully constrained model as for the other three less constrained models.

Chi-square difference tests suggest fit is not identical between the unconstrained model and the model where factor loadings are constrained to be equal between corresponding items. Similarly, fit become increasingly poor when intercepts are partially constrained and when intercepts are fully constrained. The decrement in fit was notably worse between the model with intercepts partially constrained, where intercepts of items where there was a main effect of repeated assessment on mean item values were not constrained, and the model with intercepts of all corresponding items constrained (Difference test $\chi^2(10)=499.15$) compared to the decrement in fit between the model where factor loadings were constrained and the model where intercepts were partially constrained (Difference test $\chi^2(10)=62.67$).

4.7 Research Question 3 – Synthesis of RQ1 and RQ2 Findings

> Do findings for Research Question 1 and Research Question 2 vary as a function of the way depressive symptomology is considered (i.e., assessing composite scores, assessing CES-D categorization, or at the item level)?

Figures 6 and 7 depict which terms in models of CES-D composite scores, categorizations, subscale scores, and item responses were statistically significant. Figures 6 and 7 present an overall visual summary of decisions based on observed *p*-values from the models. As noted previously, alpha was .05 for the composite score model, .025 for the categorization models, .0125 for the subscale models, and .01 for the item models. These figures display patterns across the models for visual inspection.

In Figures 6 and 7, each row represents a LMM or GEE model, with the dependent variable of each model listed in the first column. Possible terms for inclusion in the model are listed in the column headings. Each cell represents a given term (column heading) in a model for a particular dependent variable (row heading). Note that not every term in the column headings was included in each model, as these are the results from the final models after non-significant higher order interactions were dropped from the model. Figure 4 displays terms that were included in each model. Cells are shaded if the *p*-value for the represented term was less than the alpha set for a given model. Terms that are shaded with no pattern indicate the direction of the observed effect of the term was in the same direction each model. For the terms in the format column, horizontal lines in shaded cells indicate an effect in one direction (higher scores for one-item per screen format compared to multiple items per screen format) and vertical lines in shaded cells represent an effect in the opposite direction (higher scores for the multiple items per screen format compared to the one item per screen format).

Findings for research questions 1 and 2 are consistent for composite scores and both approaches to categorizing participants based on composite scores. Because the categorizations are based on composite scores, one could expect the findings to be relatively consistent; and indeed, it is interesting to note that results were consistent when two different sets of cutoff scores were used for both repeated assessment and screen format. Consistent with the results for the composite score analyses, using both categorization approaches, repeated assessment did impact the likelihood of participants being categorized into a particular depressive symptomology category, while screen format did not.

In comparing subscale results to categorization and composite score results, results regarding repeated assessment are consistent; however, results regarding format are mixed. With regard to repeated assessment, consistent with composite score and categorization results, participants reported lower subscale scores for each of the four subscales on the second assessment compared to the first assessment, where lower scores indicate lower reported depressive symptomology.

Comparing results regarding screen format, two of the four subscales, somatic symptoms and depressed affect, showed an effect of format, however no format effect was observed when considering composite scores or categorizations. For the somatic subscale, higher levels of depressive symptomology were reported on the one-item per screen format compared to the multiple items per screen format, while on the positive affect subscale higher levels of depressive symptomology were reported on the multiple items per screen format. Likely because of the opposite direction of the effect on each of these subscales, when composite scale scores are computed no effect of format was detected. As the effect was not apparent in the CES-D composite score, it is unlikely an effect would be present in categorizations based on CES-D composite scores.

Focusing on the item-level results regarding the effect of repeated assessment, participants reported lower levels of depressive symptomology on the second assessment compared to the first for eight of twenty items. Additionally, for two other items at least some participants reported lower levels of depressive symptomology on the second assessment compared to the first (e.g., for item 6, three of the four language fluency groups reported lower depressive symptomology on the second assessment).

As an effect of repeated assessment was found on all four subscales, one might expect at least one item from each subscale to also show an effect of repeated assessment. The ten items that show an effect of repeated assessment represent only three of the four subscales of the CES-D: somatic symptoms (5 of 7 items), depressed affect (4 of 7 items), and interpersonal problems (1 of 2 items). In the item analyses, none of the items from the positive affect subscale show an effect of repeated assessment. However, scores on the positive subscale are lower on the second assessment; this effect is likely driven by items 8 (Pos) and 16 (Pos), which although not statistically significant, did display a trend towards lower item responses on the second assessment (*p*=.037 and *p*=.032, respectively).

Regarding screen format, results across subscale and item responses are more mixed. As expected with no effect of screen format on composite CES-D scores and two of the subscales identified, there is no effect of format on reported CES-D scores for the majority of items. Only two of the twenty items (items 12 (Pos) and 16 (Pos)) showed differences in item responses depending on screen format. Both of these items are items on the positive affect subscale. For both of these items, participants provided higher responses when the items were presented in the multiple items per screen format, consistent with the direction of the format effect for the positive affect subscale. Although there was an effect of format on somatic symptoms subscale scores, no items from the somatic symptoms scale showed a difference in item responses depending on screen format. Interestingly, a number of items, although not statistically different, did show a trend towards significance, with *p*-values between .01 and .05. These items included the other two items from the positive affect subscale, with higher item responses on the multiple items per screen format, consistent with the results from other items on the subscale.

4.8 Research Question 4 – Effect of Select Item Characteristics on Response Change

> Are characteristics of the item (e.g., readability) predictive of participants' changes in item responses?

4.8.1 Coding of CES-D items

In addition to subscale membership, five characteristics of CES-D items were identified for consideration by looking over the twenty CES-D items for possible differences in the types of items on the scale. These characteristics included the level of reading difficulty of the item, whether the item was reverse coded referred to others, asked about perceptions of others, and asked about feelings or behaviours.

Reading difficulty was assessed using the Flesch-Kincaid grade level. Four items on the CES-D are reverse coded. For the three additional item characteristics considered, two raters coded each CES-D item according to a set coding criteria, which is included in Appendix A. Percent agreement between the two raters was 100% for the reference to others coding, 95% for the perceptions of others coding, and 85% for the behaviour/feeling coding. Table 45 presents the coding for CES-D items for each of these three characteristics, as well as Flesch-Kincaid grade level, whether the item was reverse coded, and subscale information for the items. As evident Table 45, two of the item characteristics identified, reverse coded and asking about perceptions of others, directly correspond to subscales on the CES-D. The reverse coded items include the four items that make up the positive subscale. The two items identified as referring to perceptions of others are the two items that make up the interpersonal subscale.

4.8.2 Difference in item scores

A LMM with restricted maximum likelihood estimation was used to test for differences in item responses between the first and second assessments depending on item characteristics. Using BIC fit values, a heterogeneous autoregressive one covariance matrix was selected as the repeated covariance matrix type. The model included the main effect of each of the five item characteristics, referring to others, asking about perceptions of others, asking about feelings or behaviours, the level of reading difficulty of the item, whether the item was reverse coded, as well as the screen format presented first to the participant. Because two of the identified characteristics corresponded directly to CES-D subscales, an additional indicator was added to the model to distinguish between items on the somatic subscale and other items, rather than including subscale as a factor in the model. The dependent variable was the difference in item response between the first and second assessments. Positive difference scores indicate a higher item response on the first assessment compared to the second. The order participants completed the two screen formats did not impact the difference in item responses between the two assessments ($F_{(1,\ 6187)}$=.45, p=.501; NCP 90% CI=.00-5.30).

Whether the item asked about perceptions of others ($F_{(1,\ 2901)}$=4.77, p=.029; NCP 90% CI=25-14.66), whether the item asked about a behaviour or feeling ($F_{(1,\ 4002)}$=28.96, p<.001; NCP 90% CI=13.94-49.41), whether the item was reverse coded ($F_{(1,\ 6142)}$=9.27, p=.002; NCP 90% CI=1.96-22.00), and whether the item was on the somatic subscale ($F_{(1,\ 4446)}$=11.85, p=.001; NCP 90% CI=3.23-25.89) were associated with the difference in item responses between the two assessments. Whether the item referred to others ($F_{(1,\ 5328)}$=.798, p=.372; NCP 90% CI=.00-6.43) and reading difficulty of the item ($F_{(1,\ 5296)}$=2.39, p=.122; NCP 90% CI=.00-3.54), were not associated with differences in item responses between the two assessments.

Table 46 provides the estimated marginal mean difference score between responses on the first and second assessments for items in each of the five identified categorical item characteristics. The difference between participant responses on the first assessment compared to the second assessment was larger for items coded as not asking about the perceptions of others compared to items coded as asking about perceptions of others. The difference between participant responses on the first assessment compared to the second assessment was larger for items coded as about a feeling compared to for items coded as about a behaviour. The difference between participant responses on the first assessment compared to the second was also larger for items that were not reverse coded than for items that were reverse coded and for items that were on the somatic subscale than for items that were not on the somatic subscale.

4.8.3 Change or no change in item responses

GEE analysis with a binary logistic model was used to test for differences in likelihood of participants providing a different response on the first compared to the second assessment depending on characteristics the items. QIC fit values were used to select the autoregressive one working correlation matrix. The model included the main effect of each of the item characteristics, referring to others, asking about perceptions of others, asking about feelings or behaviours, the level of reading difficulty of the item, whether the item was reverse coded as well, the somatic subscale indicator variable, as well as the screen format presented first to the participant. The dependent variable was a binary variable indicating whether a participants' response to a given item was identical or different on the first and second assessment. The order participants completed the two screen formats did not impact the difference in item responses between the two assessments (Wald $\chi^2(1)$=2.55, *p*=.110; NCP 90% CI=.00-10.51).

Reading difficulty of the item (Wald $\chi^2(1)$=8.51, *p*=.004; NCP 90% CI=1.62-20.81), whether the item referred to others (Wald $\chi^2(1)$=9.84, *p*=.002),whether the asked about perceptions of others (Wald $\chi^2(1)$=39.71, *p*<.001), whether the item asked about a behaviour or feeling (Wald $\chi^2(1)$=49.50, *p*<.001), whether the item was reverse coded (Wald $\chi^2(1)$=23.86, *p*<.001), and whether the item was on the somatic subscale (Wald

$\chi^2(1)=100.41$, $p<.001$) were associated with the difference in item responses between the two assessments. Table 46 presents 90% confidence intervals of NCPs for the effect of each coded item characteristic.

Table 46 presents the proportion of participants who provided different responses on the first assessment compared to the second assessment for the five identified categorical item characteristics. The proportion of participants who provided a different answer on the first assessment compared to the second assessment was higher for items that did not explicitly refer to others compared to items that did refer to others. The proportion of participants who provided a different answer on the first assessment compared to the second assessment was higher for items that ask about a feeling compared to items that ask about a behaviour. The proportion of participants who provided a different answer on the first assessment compared to the second assessment was higher for items that were reverse coded compared to items that were not reverse coded. The proportion of participants who provided a different answer on the first assessment compared to the second assessment was higher for items that were on the somatic subscale compared to items that were not on the somatic subscale. Higher Flesch-Kincaid grade level scores were associated with a higher likelihood of changes in responses between the first and second assessment. The mean Flesch-Kincaid grade level for items where responses were inconsistent was 2.59, while the mean Flesch-Kincaid grade level for items where responses were consistent was 2.35, where lower scores indicate items were rated as easier to read. The proportion of participants who provided a different answer on the first assessment compared to the second assessment was higher for items that were reverse coded compared to items that were not reverse coded. This is different from when difference in response values are considered, where items that were not reverse coded had a larger mean difference in response values compared to items that were reverse coded. While responses to items that were reverse coded were less likely to be consistent than items that were not reverse coded, the item responses did not differ in a consistent way for items that were reverse coded. This is in contrast to items that were not reverse coded, where item responses were more likely to be the same on the two assessments, however item responses more likely to vary in a consistent direction leading a higher mean difference score compared to reverse coded items.

5. Discussion

The present study used a repeated measures design to study the effect of screen format and repeated assessment on participant responses to a twenty item measure of depressive symptomology over the past week. The measure of depressive symptomology was considered in a number of ways including composite scale scores, subscale scores, item responses, categorizations based on scale scores, and test structure. In addition, the association of sex and self-reported English fluency with responses to the measure of depressive symptomology was considered.

The design of the study, using a repeated measures design where participants complete the measure twice in a short timeframe, addresses specific questions not addressed in previous literature. For example, do scores change with repeated assessment in the absence of an opportunity to change on the assessed construct? Additionally, by looking at the impact of repeated assessment and format at the categorization, subscale, and item level, rather than just the scale score level, this study permitted a fuller understanding of the relationship between these effects and participant responses exploring whether there was a differential pattern of effects at different levels of analysis. Furthermore, characteristics of the items were identified to try to determine whether characteristics of items were associated with participants providing different responses to items on the two assessments, potentially leading to a more clear understanding of the process involved in response consistency/inconsistency.

In the current study, the findings regarding the effect of screen format (one item per screen versus multiple items per screen) on responses to the depressive symptomology measure were mixed. There was no effect of format at the scale score and categorization level; however an effect of format was present for some subscales and items, but not others. Consistent with previous research (e.g., Arrindell, 2001), the effect of repeated assessment was present with participants reporting lower levels of depressive symptomology on the second assessment compared to the first assessment when considering overall scale scores. In addition this retest effect was relatively consistent across the measure with the effect present for categorizations based on composite scores, subscale scores, and almost half of the twenty items. The effect of screen format and repeated assessment on responses to the measure of depressive symptomology was relatively consistent for males and females and people with different levels of English fluency. Consistent with previous literature (e.g., Boticello, 2009), females reported generally higher levels of depressive symptomology, although higher composite scores appeared primarily due to differences on a few items. People who were less fluent in English also tended to report higher levels of depressive symptomology. These differences were more widespread across the measure with differences associated with different levels of English fluency on more than half of the items on the measure.

While the absolute magnitude of the effects in the study were relatively small (e.g., 1 point difference on scale score between the first and second assessment; approximately a 4 point difference between the most fluent in English and the least fluent in English groups), the study was powered to detect small effects. The results from changes in categorizations demonstrate how relatively small changes in mean score can change the way scores are interpreted with potentially important consequences.

5.1 Format

The effect of screen format on responses to the measure of depressive symptoms was mixed. Evaluating the effect on overall scale scores, no differences due to format were identified. Similarly, evidence for measurement invariance across the two formats was provided. The only exception was on the second assessment, strong factorial invariance was not found. However, when two items were dropped, strong factorial invariance was supported. This indicates on the second assessment, measurement invariance was present for the majority of the items on the measure. Measurement invariance on a measure of depressive symptomology is consistent with results of Thorndike et al. (2009) who found measurement invariance between a one item per screen format and multiple items per screen format for two other measures of depressive symptomology, the Beck Depression Inventory and the Montgomery-Asberg Depression Rating Scale. For the two items that were dropped on the second assessment when measurement invariance was identified, no differences in mean item response were present, indicating the effect of format that was observed on items 12 and 16 was likely not due to differences in the construct measured by the items.

As computerized versions of paper-and-pencil measures become more common, a range of new formats that were not practical with paper-and-pencil measures become options. For example, presenting a single item on a screen is an easily available option on computers. The measurement invariance between these new formats and formats that are structured more similarly to paper-and-pencil formats (e.g., multiple items per screen) and also the effect of these new formats on responses to questionnaire measures is key in adapting paper-and-pencil measures to electronic forms.

Inconsistent results between the mean overall composite and subscale scores highlight the importance of evaluating effects on the subscale level. While there was no effect of screen format present on scale scores, there was an effect on two of the subscales. However, likely because of the different direction of the effect on the subscales, no effect of screen format was detected at the scale score level.

The positive affect subscale was one of the subscales where screen format had an effect on subscale scores. The two items where effects of screen format were present were from the positive subscale. For the two additional items on the subscale, while not statistically different, there was a trend in the same direction. For each of these items the direction was that higher item responses were present on the multiple items per screen format. Because these items are reverse coded, if a participant with moderately low levels of depressive symptomology were to complete the measure without attending to the items and simply click down the list of items without noting the reverse coding, the participant would provide a response to the reverse coded items suggesting higher levels of depressive symptomology compared to their responses to the other items. The direction of the effect of format on subscale and item scores suggests that some participants may be more likely to respond down a list of electronic items without attending to reverse coded items compared to when the items are presented as one item per screen. This would be consistent with findings by Swartz et al. (2007) in a mode effect study, where the positive affect subscale was found to be more discriminating in the PDA mode where items were presented one item at a time compared to the paper-and-pencil mode. Swartz et al. suggested that people may be more likely to attend to each item in the PDA mode where items were presented one item at a time on the PDA device. While in the Swartz et al. study it is not possible to determine whether the positive affect subscale was more discriminating due to mode effects or because of the format (one item versus multiple) the present study suggests there is an effect of presenting one versus multiple items per screen on responses for the positive subscale.

5.2 Repeated Assessment

The CES-D was administered twice in immediate succession within a single testing session to evaluate changes in responses to a measure of depressive symptomology across assessments when there was no opportunity for real behavioural change to occur. Although correlations between timepoints were high (r=.96), mean differences were present in CES-D scores across the assessments on both scale and subscale scores and on ten items. Higher levels of depressive symptomology were reported on the first assessment compared to the second assessment. Further, the impact of these changes in reported CES-D scores on the categorization of participants based on CES-D scores was demonstrated. A higher proportion of participants were categorized in the “high depressive symptomology” group on the first assessment (.404) compared to the proportion categorized as “high depressive symptomology” on the second assessment (.355). Tests of measurement invariance also suggested that there were differences in the measurement model between the first and second assessment, particularly when the intercepts of items where a mean difference was present were fixed as equal across assessments.

These findings are consistent with previous research suggesting when measures of negative mood are completed more than once, decreased levels of negative mood will be reported at follow-up timepoints (Ormel, Koeter, & van den Brink, 1989; Sharpe & Gilbert, 1998; Arrindell, 2001).

Because the items asked about behaviour over the past week and the CES-D was completed twice within the first twenty minutes of the same testing session, a fair interpretation would be that differences in scores between assessments do not reflect a real change in level of depressive symptomology over the past week, but rather a measurement error. The presence of this retest effect when there was no opportunity for real changes in behaviour to have occurred suggests that the retest effect detected in previous research is at least in part due to a measurement artefact and not due to behavioural change caused by completing a measure or participating in a study as a wait-list control. This is consistent with Hatzenbuehler et al. (1983) and Swartz et al. (2007) who also observed lower reported levels of depressive symptomology on a second assessment given within a relatively short time period.

The present study and previous research on this retest effect highlight the importance of control groups and randomized controlled trials to ensure that mean differences in measures of negative mood which occur due to repeated assessment are not identified as a treatment effect. Additionally, results indicate caution should be taken when interpreting small differences in reported negative mood over multiple assessments using self-report questionnaires, particularly between the first and second assessment.

Additionally, these findings highlight potential challenges in studying the natural course of certain constructs, such as depressive symptomology, over time. For example, in a meta-analysis of the course of untreated depression, over a period ranging from two to twenty weeks, among people in wait-list control groups, reported depressive symptomology scores were 10%-15% lower in follow-up assessments without treatment (Posternak & Miller, 2001). In this study, over the course of twenty minutes, reported depressive symptomology scores were on average 7% lower on the second assessment. The results highlight the challenge in interpreting these types of results as entirely driven by actual changes in levels of depressive symptomology over time. Rather, reported decreases in depressive symptomology over extended periods of time in the absence of treatment may represent some combination of real change in level of depressive symptomology and this retest effect.

Arrindell (2001) describes a number of explanations that have been put forward to describe this retest effect. These explanations fall broadly into two categories 1) explanations of the retest effect as an artefact and 2) explanations of the retest effect as a change in the measured construct. These results provide evidence against explanations of the retest effect as an actual change in the construct measured because a change in mean reported scores was present in the absence of an opportunity for change in the construct measured. These results provide support for explanations of the retest effect as an artefact because changes in mean scores were present when there was no theoretical reason for a change in scores on the measure and no opportunity for a change in the construct measured to occur.

An interesting consideration related to this retest effect is whether it is caused by participants reporting increased levels of functioning on items or reporting decreased levels of negative functioning on items at follow-up timepoints. Reporting more positive functioning or reporting less negative functioning has the same impact participant scores, although the processes in play that would lead to the retest effect would be different. It is unclear what about completing measures of negative mood gives rise to this retest effect that is not occurring when completing measures of more positive states (cf., Arrindell, 2001; Sharpe & Gilbert, 1998). In the present study, lower depressive symptomology was reported on the positive affect subscale, which because of the item content and reverse coding means reporting higher levels of the positive statement. However, in the item analysis there was no effect of repeated assessment for any of the items on the positive subscale. Additionally, the confidence intervals for the non-centrality parameters for the depressed affect and somatic symptoms subscale overlap suggesting the magnitude of effect for these two subscales may be similar; in contrast, the confidence interval for the non-centrality parameter from the positive affect subscale and interpersonal problems subscale are lower and do not overlap with the other two subscales. The absence of an effect of repeated assessment detected on any of the positive affect items and lower NCP confidence intervals suggest the effect of repeated assessment was smaller for the positive affect subscale than the depressed affect and somatic symptoms subscales.

While participants did report increased positive functioning by endorsing positive statements as occurring more frequently in the past week on the second assessment compared to the first, the reported increased functioning on the positive subscale items appears less than the reported decrease in negative functioning participants reported by endorsing negative statements as occurring less frequently over the past week on the second assessment compared to the first assessment. An effect of repeated assessment on positive subscale is in contrast to findings from both Arrindell (2001) and Sharpe and Gilbert (1998). Both Arrindell and Sharpe and Gilbert did not find retest effects for positive states, including on positive subscales on measures of mood (e.g., no effect of vigour subscale on the Profile of Mood States, but an effect was present for the other five subscales of negative mood in Sharpe and Gilbert's study).

More generally, these results highlight how much we do not understand about the way people respond to a questionnaire. It is unclear whether the discrepancy between responses when completing a measure more than once occurs due to different interpretations of the questions, remembering and forming a judgement differently or selecting a different response option based on that judgement. The comparison of findings at the composite, subscale, and item level suggests that certain items are driving the retest effect on the CES-D more than others.

5.3 Item Characteristics

Item characteristics have been considered in a variety of studies (e.g., Hubley, Wu, & Zumbo, 2009), however few studies have examined their role in response consistency. In the present study, results were generally similar both when considering whether responses to an item were the same on both occasions and when considering the difference scores between the assessments. Items that were identified as reverse coded, asking about feelings, and items on the somatic subscale had higher change scores between the first and second assessment and had a higher likelihood of participants providing different responses between the two assessments. In addition, participants were more likely to provide different responses to the same item for items identified as asking about perceptions of others. Easier reading level for an item, using Flesch-Kincaid grade level scores, was associated with decreased likelihood of changing responses between the two assessments. However, reading level of the item was not associated with the mean difference scores between the two assessments, suggesting that the effect of reading difficulty may not have had an impact on responses in a consistent direction.

If participants are struggling to understand items one might expect the responses between the two assessments to vary without a consistent direction. It should be noted that participants were enrolled at a university where English is the language of instruction, with an English entry requirement. Because of the way readability was considered, length of the item was closely tied to readability scores. Another explanation is that participants were less likely to fully attend to or read longer items completely, leading to an increased likelihood of reporting scores that are different on the two assessments, but not necessarily different in a consistent direction between participants.

Findings on reverse coding, somatic subscale membership, and asking about perceptions of others, which corresponds to items on the interpersonal subscale, are consistent with the results of the repeated assessment analysis at the item level in research question 2, where an effect of repeated assessment was present for each subscale. The proportion of participants who provided a different answer between the two assessments was higher for items that were reverse coded compared to items that were not reverse coded; however the mean difference value was lower for items that were reverse coded than for items that were not reverse coded. This suggests that items that participants were more likely to provide responses that were different to items that were reverse coded, but that the differences were less likely to be in a consistent direction. Because the order formats were presented to participants was counterbalanced, this finding would be consistent with participants “straight-lining” and not noting reverse coded items in the multiple items per screen format.

5.4 Sex

Consistent with previous research, females reported higher levels of depressive symptomology compared to males (e.g., Boticello, 2009; Hankin & Abramson, 200; Van de Velde, Bracke & Levecque, 2010). Differences in item responses were found on the items asking about crying, feeling sad, feeling fearful, and being bothered, with item responses for women indicating experiencing these things more frequently during the past week. Three of these items are items from the depressed affect subscale and one from the somatic symptoms subscale.

Interpreting these higher item responses in females as indicative of higher levels of depressive symptomology is problematic in this sample because measurement invariance was not found between males and females. Previous research on the CES-D has demonstrated differential item functioning between males and females of the crying item (e.g., Gelin & Zumbo, 2003; Maller & Berkamn, 2000), meaning women with a particular level of depressive symptomology are more likely to score higher on the crying item than men with the same level of depressive symptomology. Verhoevan, Sawyer, and Spence (2012) used the same procedure to test for measurement invariance between adolescent males and females and also found that factor loadings were not the same between males and females. However, in their study, the crying item, shake off the blues item, and people were unfriendly item had different factor loadings between males and females on the first or second assessment.

In the present study, while invariance was not present between males and females, the four factor model did fit for both males and females. Particularly when the crying item (item 17) and the fearful item (item 10) were dropped from the model (RMSEA=.042 for each group). Similarly while the measurement model was not invariant between the two groups when these two items were dropped, the change in model fit was notably lower when constraints for weak factorial invariance and strong factorial invariance were added compared to when the two items were included in the model. This suggests that the crying item and fearful items may be particularly problematic in terms of comparing between males and females in this sample.

5.5 English Fluency

Level of reported English fluency was associated with reported depressive symptomology. The effect was in the direction that participants with lower levels of reported English fluency reported higher levels of depressive symptomology. This is consistent with the association between English language fluency and depressive symptomology reported in a number of populations in the US and Canada (e.g., Southeast Asian refugees in Canada (2001); immigrant Chinese adolescents in Canada, Lee and Chen (2000); international students in the US, Dao, Lee and Chang (2007)). The effect of English language fluency was present on the depressed affect subscale and positive affects subscales.

Compared to the effect of sex differences in mean reported item values, English fluency had an impact on mean reported item values on more items (11 vs 4). Reported English fluency was associated with differences on items from three of the four subscales (somatic symptoms, positive affect, interpersonal problems). Follow-up tests indicated that mean item values and reported depressive symptomology on scale and subscale scores tended to differ between participants who were most fluent in English and other language fluency groups, and between participants who were the least fluent in English and participants in other reported English fluency groups.

The test of measurement invariance indicated that the measurement model is not identical for people who were very fluent in English and participants who did not report being very fluent in English. The four factor model demonstrated fair fit for both groups, however the parameters in the model were not identical. Li and Hicks (2010) found a cultural response bias among Chinese American women on positively worded items on the CES-D (the positive subscale), where Chinese American participants who chose to complete interviews in Chinese were less likely to select highly positive responses compared to Chinese Americans who spoke English or who selected to complete the interview in English. It is possible that cultural issues around endorsing highly positive items about oneself could lead to differences in item response depending on English fluency, if English fluency is acting as a loose proxy for cultural differences or acculturation (77% of participants who did not identify as very fluent in English identified as Asian). However, in the Li and Hicks paper, cultural response bias was found between participants who completed the measure in English and those who completed the measure in Chinese, so language of the form may be the source of variation, by contrast all participants in the present study completed the measure in English and were registered at a university where English is the language of instruction, and have satisfied minimum English language proficiency requirements.

5.6 Scoring Issues

In the present study results indicate that the CES-D was not invariant between males and females and also not invariant between participants who were very fluent or English as a first language and those who were not. These results highlight that fitting the model may be not enough to establish the measurement structure if between group comparisons are planned. For example, in the present study the four factor model fit for each of the groups, however measurement invariance was still not present between groups. As one of the goals of the study was to evaluate the effect of different conditions on different scoring approaches to the CES-D, standard scoring was used in subsequent analyses despite issues of measurement invariance.

5.7 Limitations

This study was conducted on undergraduate students and thus the generalizability of the findings from this study is an issue. However, even if the findings are not generalizable to the general population, the issue of mental health and depressive symptomology in undergraduate students is relevant in and of itself (e.g., Lunau, 2012; Storrie, Ehern, & Tucker, 2010).

One specific challenge with this population in terms of screen format effects is the relative computer competency of the population. As the sample was on average relatively young (20 years) and would likely have exposure to computers through coursework, they may be better with computers and reading off of a screen than other parts of the general population. It is possible that for people who are less comfortable with computers and reading from a screen, format may have a different effect. For example, Swartz et al. (2007) found an interaction between questionnaire mode (paper-and-pencil and PDA) and education, where as education level increased the difference between scores on paper-and-pencil and PDA modes decreased. Replication would be necessary to evaluate whether the results generalize to other groups.

In terms of the retest effect, generalizability may be an issue as well. However, the present study of immediate test-retest effects on depressive symptomology is building on the observation of test-retest effects in studies based on undergraduate students, clinical populations, as well as community samples over longer periods of time. Although it is an empirical question that only research can address, the findings of immediate test-retest effects observed in the current study may generalize to a broader sample.

A second limitation is a challenge that is present when factors that impact the measurement process or measurement error are studied. A target measure must be selected to study the measurement process. Replication across measures is necessary to generalize the process beyond the particular measure used, however multiple measures demonstrating a similar effect, such as the retest effect which has been observed on multiple measures of negative mood, lead to increased confidence in the generalizability to other measures and the presence of a more general measurement process.

5.8 Future Directions

Expanding the present study to address some of the issues of generalizability would be useful. For example, using different measures or sampling different populations with less computer experience. In terms of retest, in addition to consideration of other item characteristics, future studies could further compare positive and negative items presented together to further understand the process and why an impact of repeated assessment was present on positively worded items on a measure of negative mood, but generally this effect of repeated assessment has not been found on measures of positive moods.

In terms of format, testing screen formats which have shown to have an impact in paper-and-pencil modes to determine if there are similar effects when electronic modes are used. This could suggest that the processes leading to reported scores are similar between modes. For example, Toepoel, Das, and van Soest (2008) reported participants tended not to select response options that had negative values as category labels. If a similar effect is present in paper-and-pencil responses, a similar process could occur to produce these similar effects.

With regard to item characteristics, the current study looked at structural as well as content features of items. Further examination of structural features and content features may further inform understanding of response processes. For example, with regard to item readability, the current study used a global index of readability; future studies examining items with regard to specific aspects of readability (e.g., word length) will be informative.

5.9 Recommendations

Although previous literature has demonstrated that people may report higher functioning on a follow-up assessment compared to the first assessment in certain measures of negative mood when there is no theoretical reason for the reported change (e.g., Arrindell, 2001; Sharpe & Gilbert, 1998; Ahava et al., 1998), the current study demonstrates participants report higher functioning on a second assessment of depressive symptomology in the absence of an opportunity for real change in depressive symptomology.

Previous research on the effect of screen format have found consistent factor structure on certain measures of depressive symptomology (Thorndike et al., 2009); the present study provides evidence for measurement invariance between single item and multi item screen formats on the CES-D. However, some trends in the item level responses suggested reverse coded items were less likely to be noted by participants in the multiple items per screen format compared to the single item per screen format.

Based on integration of previous literature and the current study, the following recommendations are made. First, in experimental longitudinal studies of negative mood, wait-list control conditions are recommended, so changes in reported negative mood due to measurement can be accounted for. Next, in correlational longitudinal studies of negative mood, caution in terms of interpreting small changes in negative mood is recommended, particularly in the absence of an a priori theoretical reason to expect change in reported negative mood. In general when presenting the CES-D electronically, multiple items per screen and single item per screen formats could be used, however in the multiple item form, identifying participants who likely did not note reverse coded items (i.e., responded in a straight line down the screen) may improve data quality. Finally, careful consideration of item characteristics to minimize measurement artefacts is recommended; such as reverse coding and item readability. And although the current study did not show interaction effects of sex and language fluency in the research questions examined, the literature and this study highlight that consideration of participant characteristics in models of responses to questionnaires is essential.

References

Ahava, G. W., Iannone, C., Grebstein, L., & Schirling, J. (1998). Is the Beck Depression Inventory reliable over time? An evaluation of multiple test-retest reliability in a nonclinical college student sample. *Journal of Personality Assessment, 70(2),* 222-231.

Akhtar-Danesh, N., Landeen, J. (2007). Relation between depression and sociodemographic factors. *International Journal of Mental Health Systems, 1(4).* Retrieved October 26, 2010, from http://www.ijmhs.com/content/1/1/4

Arrindell, W. A. (2001). Changes in waiting-list patients over time: data on some commonly-used measures. Beware! *Behavior Research and Therapy, 39,* 1227-1247.

Babbie, E. (1990). *Survey Research Methods* (2nd ed.). Belmont, CA: Wadsworth.

Beiser, M. N., & Hou, F. (2001). Language acquisition, unemployment and depressive disorder among Southeast Asian refugees: A 10-year study. *Social Science & Medicine, 53*, 1321-1334.

Biemer, P. P., Groves, R. Lyberg, L. E., Mathiowez, N., & Sudman, S. eds. (1991). *Measurement errors in surveys.* New York: John Wiley & Sons, Inc.

Botticello, A. (2009). A multilevel analysis of gender differences in psychological distress over time. *Journal of Research on Adolescence, 19(2),* 217-247.

Brown, C., Schale, C. L., Nilsson, J. E. (2010). Vietnamese immigrant and refugee women's mental health: An examination of age of arrival, length of stay, income, and English language proficiency. *Journal of Multicultural Counseling and Development, 38(2),* 66-76.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (eds.), *Testing Structural Equation Models*. (pp. 136-162). Newbury Park, CA: Sage.

Christian, L. M., & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly, 68,* 58-81.

Cole, S. R., Kawachi, I., Maller, S. J., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE study. *Journal of Clinical Epidemiology, 53,* 285-289.

Couper, M. P., Traugott, M. W., Lamais, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly, 65,* 230-253.

Culbertson, F. M. (1997). Depression and gender: An international review. *American Psychologist, 52(1),* 25-31.

Dao, T. K., Lee, D., & Chang, H. L. (2007). Acculturation level, perceived English fluency, perceived social support level, and depression among Taiwanese international students. *College Student Journal, 41(2),* 287-295.

Deardoff, W. W., & Funabiki, D., (1985). A diagnostic caution in screening for depressed college students. *Cognitive Therapy and Research, 9(3),* 277-284.

Flanagan, D. P., & Ortiz, S. O. (2001). *Essentials of cross-battery assessment*. New York: Wiley.

Fouladi, R. T., McCarthy, C. J., & Moller, N. P. (2002). Paper-and-pencil or online? Evaluating mode effects on measures of emotional functioning and attachment. *Assessment, 9(2),* 204-215.

French, D. J. (2012). A simple measure with complex determinants: investigation of the correlates of self-rated health in older men and women from three continents. *BMC Public Health, 12(1),* 649.

Garber, J., Clarke, G. N., Weersing, V. R., Beardslee, W. R., Brent, D. A., Gladstone, T. R., DeBar, L. L., Lynch, F. L., D'Angelo, E., Hollon, S. D., Shamseddeen, W., & Iyengar, S. (2009). Prevention of depression in at-risk adolescents: A randomized controlled trial. *Journal of the American Medical Association, 301(21),* 2215-2224.

Gelin, M. A., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression scale. *Educational and Psychological Measurement, 63(1),* 65-74

Gibbons, J. L., Zellner, J. A., & Rudek, D. J. (1999). Effects of language and meaningfulness on the use of extreme response style by Spanish-English bilinguals. *Cross-Cultural Research, 33,* 369-381.

Grice, H. P. (1975). Logic and Conversation In P. Cole and J. L Morgan (eds.) *Syntax and Semantics: Vol. 3: Speech Acts,* New York, Academic Press, 41-58.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74(5),* 849-879.

Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences, 44,* 932-943.

Hankin, B. L., & Abramson, L. Y. (2001). Development of gender differences in depression: An elaborated cognitive vulnerability-transactional stress theory. *Psychological Bulletin, 127(6),* 773-796.

Hartley, J., & Betts, L. R. (2009). Four layouts and a finding: the effects of changes in the order of the verbal labels and numerical values on Likert-type scales. *International Journal of Social Research Methodology, 13(1),* 17-27.

Harzing, A. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management, 6, 243-266.*

Hatzenbuehler, L. C., Parpal, M., & Matthews, L. (1983). Classifying college students as depressed or nondepressed using the Beck Depression Inventory: An empirical analysis. *Journal of Consulting and Clinical Psychology, 51(3),* 360-366.

Heerwegh, D., & Loosveldt, G. (2002). An evaluation of the effect of response formats on data quality in web surveys. *Social Science Computer Review, 20,* 471-484.

Hintze, J. (2008). PASS 2008. [Computer software]. Kaysville, UT: NCSS, LLC.

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

Hubley, A. M., Wu, A. D., & Zumbo, B. D. (2009). Interpreting IRT parameters: Putting psychological meat on the psychometric bone. *Presented at the Annual Meeting of the American Psychological Association,* Toronto, Canada.

Hui, C. H., Tirandis, H., C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20(3),* 296-309.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55(4),* 243-252.

Joint Committee for Guides in Metrology. (2012). *International vocabulary of basic and general terms in metrology (VIM) 3rd edition.* Retrieved November 30, 2012, from the BIPM website: http://www.bipm.org/en/publications/guides/vim.html

Kessler, R. C., McGonagle, K. A., Swartz, M., Blazer, D. G., & Nelson, C. B. (1993). Sex and depression in the National Comorbidity Survey I: Lifetime prevalence, chronicity and recurrence. *Journal of Affective Disorders, 29,* 85-96.

Kiesler, S., & Sproull, L. S. (1986). Response effects in the electronic survey. *Public Opinion Quarterly, 50(3),* 402-413.

Kroenke, K., Wu, J., Bair, M. J., Krebs, E. E., Damush, T. M., & Wu, T. (2011). Reciprocal relationship between pain and depression: A 12-month longitudinal analysis in primary care. *Journal of Pain, 12(9),* 964-973.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50,* 537-567.

Kuehner, C. (2003). Gender differences in unipolar depression: An update of epidemiological findings and possible explanations. *Acta Psychiatrica Scandinavia, 108,* 163-174.

Lange, R. Thalbourne, M. A., Houran, J., & Lester, D. Depressive response sets due to gender and culture-based differential item functioning, 33, 937-954.

Lee, B. K., & Chen, L. (2000). Cultural communication competence and psychological adjustment – A study of Chinese immigrant children's cross-cultural adaptation in Canada. *Communication Research, 27,* 764-792.

Locke, K. D., & Baik, K. (2009). Does an acquiescent response style explain why Koreans are less consistent than Americans? *Journal of Cross-Cultural Psychology, 40,* 319-323.

Longwell, T. B., & Truax, P. (2005). The differential effects of weekly, monthly, and bimonthly administrations of the beck depression inventory-II: Psychometric properties and clinical implications. *Behavior Therapy, 36(3),* 265-275.

Lunau, K. (September 5th, 2012). The mental health crisis on campus: Canadian students feel hopeless, depressed, even suicidal. *Macleans Magazine.* Canada. Accessed October 6, 2012, from http://oncampus. macleans.ca/education/2012/09/05/the-mental-health-crisis-on-campus/

Mahon-Haft, T. A., & Dillman, D. A. (2010). Does visual appeal matter? Effects of web survey aesthetics on survey quality. *Survey Research Methods, 4(1),* 43-59.

Maier, W., Gansicke, M., Gater, R., Rezaki, M., Tiemens, B., & Urzua, R. F. (1999). Gender differences in the prevalence of depression: A survey in primary care. *Journal of Affective Disorders, 53,* 241-252.

Mari, L. (2005). The problem of foundations of measurement. *Measurement, 38(4),* 259-266.

McCabe, S. E., Boyd, C. J., Couper, M. P., Crawford, S., & D'Arcy, H. (2002). Mode effects for collecting alcohol and other drug use data: Web and U.S. mail. *Journal of Studies on Alcohol, 63(6),* 755-761.

Mplus website (n.d.) Chi-square difference testing using the Satorra-Bentler scaled chi-square. Retrieved October 2, 2013, from http://www.statmodel.com/chidiff.shtml.

Muthén, B., & Muthén, L. (Authors, Copyright holders). (2009). Topic 1. Introductory – advanced factor analysis and structural equation modeling with continuous outcomes [Video for Mplus short courses]. (Available from Mplus website, http://statmodel2.com/course_materials.shtml)

Muthén, B., & Muthén, L. (Authors, Copyright holders). (2010). Topic 3. Introductory and intermediate growth modeling. [Video for Mplus short courses]. (Available from Mplus website, http://statmodel2.com/course_materials.shtml)

Niemi, I. (1993). Systematic error in behavioural measurement: Comparing results from interview and time budget studies. *Social Indicators Research, 30,* 229-244.

Nikolova, Y. S. (2012). Ventral striatum reactivity to reward and recent life stress interact to predict positive affect. *Biological Psychiatry, 72(2),* 157-163.

Ormel, J., Koeter, M. W., & van den Brink, W. (1989). Measuring change with the general health questionnaire. *Social Psychiatry and Psychiatric Epidemiology, 24,* 227-232.

Orth, U., Robings, R. W., Meier, L. L. (2009). Disentangling the effects of low self-esteem and stressful events on depression: Findings from three longitudinal studies. *Journal of Personality and Social Psychology, 97(2),* 307-321.

Pace, W. D., & Staton, E. W. (2005). Electronic data collection options for practice-based research networks. *Annals of Family Medicine, (3),* S21-S29.

Patten, S. B., Lavorato, D. H., Metz, L. M. (2005). Clinical correlates of CES-D depressive symptom ratings in an MS population. *General Hospital Psychiatry, 27(6),* 439-445.

Peytchev, A., Couper, M. P., McCabe, S. E., & Crawford, S. D. (2006). Web survey design: Paging versus scrolling. *Public Opinion Quarterly, 70(4),* 596-607.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1(3),* 385-401.

Radloff, L. S. (1991). The use of the Center for Epidemiologic Studies Depression Scale in adolescents and young adults. *Journal of Youth and Adolescence, 20(2),* 149-166.

Reips, U. D. (2002). Internet-based psychological experimenting: Five does and five don’ts. *Social Science Computer Review, 20(3)*, 241-249.

Roberts, R. E., Lewinsohn, P. M., & Seeley, J. R. (1991). Screening for adolescent depression: A comparison of depression scales. *Journal of the American Academy of Child & Adolescent Psychiatry, 30,* 58–66.

Rumbaut, R. G. (1994). The crucible within: Ethnic identity, self-esteem, and segmented assimilation among children of immigrants. *International Migration Review, 28,* 748-794.

Rushton, J. L., Forcier, M., & Schectman, R. M. (2002). Epidemiology of depressive symptoms in the national longitudinal study of adolescent health. *Journal of the American. Academy of Child and Adolescent Psychiatry, 41(2),* 199-205.

Satorra, A., & Bentler, P. M. (1999). A scaled difference chi-square test statistics for moment structure analysis. *Psychometrika, 66(4),* 507-514.

Scharer, L. O., Hartweg, V., Hoern, M., Graesslin, Y., Frey, C., Waler. S., et al. (2002). Electronic diary for bipolar patients. *Neuropsychobiology, 46(supp1)*, 10-12.

Schwarz, N. (1995). What respondents learn from questionnaires: The survey interview and the logic of conversation. *International Statistical Review, 63(2),* 153-177.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22(2),* 127-160.

Schwarz, N., Grayson, C. E., & Knauper, B. (1998). Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research, 10(2),* 177-183.

Shafer, A. B. (2006). Meta-anlaysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology, 62(1),* 123-146.

Shannon, D. M., & Bradshaw, C. C. (2002). A comparison of response rate, response time, and costs of mail and electronic surveys. *The Journal of Experimental Education, 70(2),* 179-192.

Sharpe, J. P., & Gilbert, D. G. (1998). Effects of repeated administration of the Beck Depression Inventory and other measures of negative mood states. *Personality and Individual Differences, 24(4),* 457-463.

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology, 35,* 50-61.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Effects of using visual design principles to group response options in web surveys. *International Journal of Internet Science, 1(1),* 6-16.

Steiger, J. (n.d.). Noncentral distribution calculator (NDC) [Computer software]. Retrieved from http://www.statpower.net/Software.html

Steiger, J. & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Muliak, & J. Steiger (Eds.), *What if there were no significance tests?* (pp.59-81). Amsterdam, Netherlands: Elsevier.

Storrie, K., Ehern, K., & Tucker, A. (2010). A systematic review: Students with mental health problems – A growing problem. *International Journal of Nursing Practice, 16,* 1-6.

Stommel, M. Given, B. A., Given, C. W., Kalaian, H. A., Schulz, R., & McCorkle, R. (1993). Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression scale (CES-D). *Psychiatry Research, 49,* 239-250.

Swartz, R. J., de Moor, C., Cook, K. F., Fouladi, R. T., Basen-Engquist, K., Eng, C. & Carmack, T. C. L. (2007) Mode effects in the Center for Epidemiologic Studies Depression (CES-D) Scale: Personal digital assistant vs. Paper and pencil administration. *Quality of Life Research, 16(5),* 803-813.

Thorndike, F. P., Carlbring, P., Smyth, F. L., Magee, J. C., Gonder-Frederick, L., Ost, L., Ritterband, L. M. (2009). Web-based measurement: Effect of completing single or multiple items per webpage. *Computers in Human Behavior, 25,* 393-401.

Thriemer, K., Ley, B. B., Ame, S. S., Deen, J. L., Pak, G. D., Chang, N. Y., et al. (2012). Clinical and epidemiological features of typhoid fever in Pemba, Zanzibar: Assessment of the performance of the WHO case definitions. *PLOS One.* Retrieved September 14, 2013, from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0051823

Toepoel, V., & Couper, M. P. (2011). Can verbal instructions counteract visual context effects in web surveys? *Public Opinion Quarterly, 75(1),* 1-18.

Toepoel, V., Das, M., & van Soest, A. (2008). Effects of design in web surveys: Comparing trained and fresh respondents. *Public Opinion Quarterly, 72(5),* 985-1007.

Toepoel, V., Das, M., & van Soest, A. (2009). Design of web questionnaires: the effects of the number of items per screen. *Field Methods, 21(2),* 200-213.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133(5),* 859-883.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order: interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly, 68,* 368-393.

Vahdaninia, M. Omidvari, S., & Montazeri, A. (2010). What do predict anxiety and depression in breast cancer patients? A follow-up study. *Social Psychiatry and Psychiatric Epidemiology, 45,* 355-361.

Van de Velde, S., Bracke, P., & Levecque. (2010). Gender differences in depression in 23 European countries. Cross-national variation in the gender gap in depression. *Social Science & Medicine, 71,* 305-313.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35,* 346-360.

Velez, P., & Ashworth, S. D. (2007). The impact of item readability on the endorsement of the midpoint response in surveys. *Survey Research Methods, 1(2),* 69-74.

Verhoeven, M., Sawyer, M. G., & Spence, S. H. (2013). The factorial invariance of the CES-D during adolescence: Are symptom profiles for depression stable across gender and time? *Journal of Adolescence, 36,* 181-190.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117(2),* 250-270.

Wade, T. J., Cairney, J., & Pevalin, D. J. (2002). Emergence of gender differences in depression during adolescence: National panel results from three countries. *Journal of the American Academy of Child Adolescent Psychiatry, 41(2),* 190-198.

Wallis, P., & Fouladi, R. T. (2007). Assessing Participant Preference in Electronic Questionnaire Format. *Poster presented at the Western Psychological Association Conference.*

Watkins. E. R., Baeyens, C. B., & Read, R. (2009). Concreteness training reduces dysphoria: Proof-of-principle for repeated cognitive bias modification in depression. *Journal of Abnormal Psychology, 118(1),* 55-64.

Weissman, M. M., Bland, R. C., Canino, G. J., Faravelli, C., Greenwald, S., Hwu, H., et al. (1996). Cross-national epidemiology of major depression and bipolar disorder. *Journal of the American Medical Association, 276,* 293-299.

Weissman, M. M., Sholomskas, D., Pottenger, M., Prusoff, B. A., & Locke, B. Z. (1977). Assessing depressive symptoms in five psychiatric populations: A validation study. *American Journal of Epidemiology, 106(3),* 203-214.

Appendix A.

Details of Identified Item Characteristics

Flesch-Kincaid Grade Level

(.39 x ASL) + (11.8 x ASW) – 15.59

where:

ASL = average sentence length (the number of words divided by the number of sentences)

ASW = average number of syllables per word (the number of syllables divided by the number of words)

Reference to others

Explicit Reference to Others: Does the item explicitly (i.e., directly mention) refer to a person/people other than the participant?

Perception of others

Asks about Perceptions of Others: Does the item explicitly (i.e., directly mention) ask about the beliefs or behaviour of others towards the respondent?

Behaviour/Feeling

Does the item ask about a behaviour the participant may or may not engage in. Items that are structured as asking about feelings about behaviours (e.g., “I felt things were going badly”) are classified in the “asks about a feeling” category.

Appendix B.

Reference Table of NCP Values and Corresponding Partial *eta*^{2}, *eta*, *f*^{2} and *f* for $\chi^2(1)$ and *F* tests with *N*=940

NCP	*Wald* χ^2		*GLM F*			
	eta^2	*eta*	*eta*^{2}	*eta*	*f*^{2}	*f*
10	.011	.103	.011	.103	.011	.011
20	.021	.146	.021	.144	.021	.021
30	.032	.179	.031	.176	.032	.031
40	.043	.206	.041	.202	.043	.041
50	.053	.231	.051	.225	.053	.051
60	.064	.253	.060	.245	.064	.060
70	.074	.273	.069	.263	.074	.069
80	.085	.292	.078	.280	.085	.078
90	.096	.309	.087	.296	.096	.087
100	.106	.326	.096	.310	.106	.096
110	.117	.342	.105	.324	.117	.105
120	.128	.357	.113	.336	.128	.113
130	.138	.372	.121	.349	.138	.121
140	.149	.386	.130	.360	.149	.130
150	.160	.399	.138	.371	.160	.138
160	.170	.413	.145	.381	.170	.145
170	.181	.425	.153	.391	.181	.153
180	.191	.438	.161	.401	.191	.161
190	.202	.450	.168	.410	.202	.168
200	.213	.461	.175	.419	.213	.175
210	.223	.473	.183	.427	.223	.183

Table 1. List of CES-D Items and Subscales

Som	1.	I was bothered by things that usually don’t bother me.
Som	2.	I did not feel like eating; my appetite was poor.
Dep	3.	I felt that I could not shake off the blues even with help from my family or friends.
Pos	4.	I felt I was just as good as other people.
Som	5.	I had trouble keeping my mind on what I was doing.
Dep	6.	I felt depressed.
Som	7.	I felt that everything I did was an effort.
Pos	8.	I felt hopeful about the future.
Dep	9.	I thought my life had been a failure.
Dep	10.	I felt fearful.
Som	11.	My sleep was restless.
Pos	12.	I was happy.
Som	13.	I talked less than usual.
Dep	14.	I felt lonely.
Int	15.	People were unfriendly.
Pos	16.	I enjoyed life.
Dep	17.	I had crying spells.
Dep	18.	I felt sad.
Int	19.	I felt that people dislike me.
Som	20.	I could not get “going.”

Table 2. CES-D Items listed by Subscale

Somatic Symptoms:

1. I was bothered by things that usually don’t bother me.
2. I did not feel like eating; my appetite was poor.
5. I had trouble keeping my mind on what I was doing.
7. I felt that everything I did was an effort.
11. My sleep was restless.
13. I talked less than usual.
20. I could not get “going.”

Depressed Affect:

3. I felt that I could not shake off the blues even with help from my family or friends.
6. I felt depressed.
9. I thought my life had been a failure.
10. I felt fearful.
14. I felt lonely.
17. I had crying spells.
18. I felt sad.

Positive Affect:

4. I felt I was just as good as other people.
8. I felt hopeful about the future.
12. I was happy.
16. I enjoyed life.

Interpersonal Problems:

15. People were unfriendly.
19. I felt that people dislike me.

Table 3. Descriptive Statistics of CES-D Scores at each Timepoint and for each Format

								Cronbach's alpha		
	Min	Max	Median	Mean	*SD*	Skew (*SE*)	Kurtosis (*SE*)	α	95% CI	
Time of assessment										
First presentation	0	47	13.84	15.21	9.08	0.84 (.08)	0.29 (.16)	.88	.87	.89
Second presentation	0	46	12	14.18	9.22	0.95 (.08)	0.61 (.16)	.89	.88	.90
Format										
One-item per screen	0	47	13	14.64	9.20	.91 (.08)	0.54 (.16)	.89	.88	.90
Multiple items per screen	0	45	13	14.75	9.14	.86 (.08)	0.34 (.16)	.88	.87	.89

Note. N = 940 for each row.

Table 4. Descriptive Statistics for each of the Four CES-D Subscales at each Timepoint

									Cronbach’s alpha		
	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}	α	95% CI	
Somatic symptoms											
First presentation	0	19	6.54	3.33	.59	.08	.38	.16	.63	.59	.67
Second presentation	0	19	6.11	3.46	.60	.08	.39	.16	.69	.66	.72
Depressed affect											
First presentation	0	20	4.07	3.96	1.20	.08	1.07	.16	.84	.83	.86
Second presentation	0	20	3.68	4.03	1.41	.08	1.71	.16	.87	.86	.89
Positive affect											
First presentation	0	12	3.76	2.85	0.53	.08	-0.45	.16	.81	.79	.83
Second presentation	0	12	3.63	2.92	0.51	.08	-0.53	.16	.85	.83	.86
Interpersonal problems											
First presentation	0	6	.85	1.13	1.49	.08	2.14	.16	.58	.53	.63
Second presentation	0	6	.76	1.14	1.69	.08	2.86	.16	.71	.68	.75

Note. N = 940 for each row. Theoretical range for each subscale was: somatic symptoms 0-21, depressed affect 0-21, positive affect 0-12, interpersonal problems 0-6.

Table 5. Descriptive Statistics for each of the Four CES-D Subscales for each Screen Format

									Cronbach's alpha		
	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}	α	95% CI	
Somatic symptoms											
One item per screen	0	19	6.38	3.37	.55	.08	.36	.16	.66	.63	.69
Multiple items	0	19	6.27	3.43	.63	.08	.39	.16	.67	.63	.70
Depressed affect											
One item per screen	0	20	3.89	4.03	1.32	.08	1.42	.16	.86	.85	.87
Multiple items	0	20	3.86	3.96	1.28	.08	1.30	.16	.86	.84	.87
Positive affect											
One item per screen	0	12	3.58	2.87	.54	.08	-.47	.16	.84	.82	.85
Multiple items	0	12	3.81	2.90	0.50	.08	-.51	.16	.82	.80	.84
Interpersonal problems											
One item per screen	0	6	.79	1.12	1.62	.08	2.66	.16	.68	.64	.50
Multiple items	0	6	.82	1.15	1.56	.08	2.32	.16	.62	.57	.66

Note. N = 940 for each row. Theoretical range for each subscale was: somatic symptoms 0-21, depressed affect 0-21, positive affect 0-12, interpersonal problems 0-6.

Table 6. Descriptive Statistics for each of the Twenty CES-D Items on the First and Second Presentation

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 1									
First presentation	939	0	3	.80	.80	.75	.08	-.02	.16
Second presentation	938	0	3	.69	.74	.86	.08	.23	.16
Item 2									
First presentation	937	0	3	.55	.77	1.30	.08	1.00	.16
Second presentation	929	0	3	.52	.74	1.36	.08	1.33	.16
Item 3									
First presentation	930	0	3	.66	.84	1.09	.08	.36	.16
Second presentation	925	0	3	.55	.75	1.25	.08	.94	.16
Item 4									
First presentation	931	0	3	.97	.96	.61	.08	-.70	.16
Second presentation	936	0	3	.95	.93	.56	.08	-.74	.16
Item 5									
First presentation	937	0	3	1.49	.88	.03	.08	-.71	.16
Second presentation	934	0	3	1.35	.88	.22	.08	-.63	.16
Item 6									
First presentation	939	0	3	.68	.86	1.12	.08	.45	.16
Second presentation	935	0	3	.60	.85	1.32	.08	.89	.16
Item 7									
First presentation	937	0	3	1.24	.91	.32	.08	-.69	.16
Second presentation	940	0	3	1.16	.93	.43	.08	-.67	.16
Item 8									
First presentation	934	0	3	1.07	.90	.41	.08	-.71	.16
Second presentation	935	0	3	1.03	.91	.45	.08	-.73	.16
Item 9									
First presentation	936	0	3	.32	.63	2.06	.08	4.00	.16
Second presentation	934	0	3	.29	.60	2.25	.08	4.93	.16
Item 10									
First presentation	939	0	3	.53	.76	1.31	.08	1.01	.16
Second presentation	936	0	3	.46	.74	1.62	.08	2.12	.16

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 11									
First presentation	939	0	3	.93	.93	.67	.08	-.52	.16
Second presentation	933	0	3	.92	.93	.71	.08	-.46	.16
Item 12									
First presentation	931	0	3	.86	.83	.63	.08	-.36	.16
Second presentation	932	0	3	.84	.85	.70	.08	-.34	.16
Item 13									
First presentation	931	0	3	.77	.81	.80	.08	-.04	.16
Second presentation	938	0	3	.69	.75	.84	.08	.17	.16
Item 14									
First presentation	935	0	3	.82	.90	.87	.08	-.14	.16
Second presentation	934	0	3	.75	.87	1.00	.08	.19	.16
Item 15									
First presentation	938	0	3	.41	.66	1.61	.08	2.20	.16
Second presentation	937	0	3	.38	.65	1.74	.08	2.74	.16
Item 16									
First presentation	935	0	3	.86	.88	.66	.08	-.53	.16
Second presentation	935	0	3	.81	.85	.71	.08	-.43	.16
Item 17									
First presentation	938	0	3	.31	.65	2.29	.08	5.01	.16
Second presentation	939	0	3	.32	.66	2.25	.08	4.80	.16
Item 18									
First presentation	934	0	3	.74	.82	.92	.08	.19	.16
Second presentation	934	0	3	.70	.83	1.05	.08	.44	.16
Item 19									
First presentation	936	0	3	.44	.68	1.54	.08	2.03	.16
Second presentation	936	0	3	.39	.65	1.70	.08	2.57	.16
Item 20									
First presentation	938	0	3	.76	.84	.95	.08	.20	.16
Second presentation	936	0	3	.77	.85	.89	.08	.01	.16

Table 7. Descriptive Statistics for each of the Twenty CES-D Items for each Screen Format

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 1									
One item per screen	938	0	3	.76	.78	.77	.08	.00	.16
Multiple items	939	0	3	.73	.77	.85	.08	.23	.16
Item 2									
One item per screen	936	0	3	.53	.74	1.37	.08	1.37	.16
Multiple items	930	0	3	.55	.77	1.30	.08	.97	.16
Item 3									
One item per screen	927	0	3	.60	.78	1.20	.08	.76	.16
Multiple items	928	0	3	.62	.81	1.15	.08	.54	.16
Item 4									
One item per screen	934	0	3	.93	.93	.61	.08	-.66	.16
Multiple items	933	0	3	.99	.96	.56	.08	-.77	.16
Item 5									
One item per screen	935	0	3	1.44	.87	.12	.08	-.64	.16
Multiple items	936	0	3	1.40	.90	.13	.08	-.74	.16
Item 6									
One item per screen	936	0	3	.64	.86	1.22	.08	.61	.16
Multiple items	938	0	3	.64	.85	1.22	.08	.69	.16
Item 7									
One item per screen	940	0	3	1.22	.93	.36	.08	-.70	.16
Multiple items	937	0	3	1.18	.92	.39	.08	-.67	.16
Item 8									
One item per screen	936	0	3	1.02	.91	.45	.08	-.75	.16
Multiple items	933	0	3	1.07	.91	.42	.08	-.70	.16
Item 9									
One item per screen	934	0	3	.30	.61	2.18	.08	4.62	.16
Multiple items	936	0	3	.30	.62	2.13	.08	4.25	.16
Item 10									
One item per screen	937	0	3	.49	.75	1.49	.08	1.54	.16
Multiple items	938	0	3	.51	.75	1.43	.08	1.49	.16

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 11									
One item per screen	932	0	3	.93	.92	.70	.08	-.42	.16
Multiple items	940	0	3	.92	.94	.69	.08	-.54	.16
Item 12									
One item per screen	935	0	3	.83	.82	.69	.08	-.24	.16
Multiple items	928	0	3	.88	.86	.64	.08	-.45	.16
Item 13									
One item per screen	937	0	3	.73	.77	.81	.08	.08	.16
Multiple items	932	0	3	.73	.79	.84	.08	.07	.16
Item 14									
One item per screen	937	0	3	.81	.91	.90	.08	-.09	.16
Multiple items	932	0	3	.77	.87	.96	.08	.12	.16
Item 15									
One item per screen	936	0	3	.39	.63	1.58	.08	2.00	.16
Multiple items	939	0	3	.40	.67	1.75	.08	2.78	.16
Item 16									
One item per screen	936	0	3	.80	.84	.70	.08	-.44	.16
Multiple items	934	0	3	.87	.90	.66	.08	-.54	.16
Item 17									
One item per screen	937	0	3	.31	.65	2.26	.08	4.87	.16
Multiple items	940	0	3	.31	.65	2.28	.08	4.94	.16
Item 18									
One item per screen	935	0	3	.74	.83	.96	.08	.26	.16
Multiple items	933	0	3	.70	.81	1.01	.08	.37	.16
Item 19									
One item per screen	937	0	3	.40	.65	1.60	.08	2.19	.16
Multiple items	935	0	3	.42	.68	1.63	.08	2.33	.16
Item 20									
One item per screen	937	0	3	.77	.85	.90	.08	.08	.16
Multiple items	937	0	3	.76	.85	.93	.08	.13	.16

Table 8. Descriptive Statistics of Composite CES-D Scores for each Screen Format on the First and Second Assessment

									Cronbach's alpha		
	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}	α	95% CI	
First presentation											
One Item Per Screen[a]	0	47	15.05	8.84	.95	.12	.63	.23	.88	.86	.89
Multiple Items[b]	0	44	15.34	9.29	.75	.11	.06	.22	.88	.86	.90
Second presentations											
One Item Per Screen[b]	0	46	14.28	9.50	.90	.11	.49	.22	.90	.89	.91
Multiple Items[a]	0	45	14.06	8.91	1.00	.12	.78	.23	.88	.86	.90

Note. [a] *n*= 470 for participants who completed one item per screen format first and multiple items per screen format second. [b] *n*= 470 for participants who completed multiple items per screen format first and one item per screen format second.

Table 9. Descriptive Statistics of Composite CES-D Scores for Males and Females

									Cronbach’s alpha		
	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}	α	95% CI	
First presentation											
Male	0	46	14.40	8.62	.95	.14	.66	.27	.87	.85	.89
Female	0	47	15.63	9.29	.78	.10	.14	.20	.89	.87	.90
Second presentations											
Male	0	43	13.27	8.84	1.036	.14	.97	.27	.89	.87	.90
Female	0	46	14.66	9.40	.90	.10	.46	.20	.90	.88	.91
One item per screen											
Male	0	46	13.84	8.80	1.032	.14	1.06	.27	.88	.86	.90
Female	0	47	15.06	9.38	.85	.10	.33	.20	.90	.88	.91
Multiple items per screen											
Male	0	43	13.83	8.69	.94	.14	.53	.27	.87	.85	.89
Female	0	45	15.23	9.33	.82	.10	.24	.20	.89	.87	.90

Note. $n_{\text{male}} = 324$; $n_{\text{female}} = 616$.

Table 10. Descriptive Statistics of Composite CES-D Scores for Participants with Different Levels of Self-rated English Fluency by Screen Format and by Assessment Time

									Cronbach's alpha		
	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}	α	95% CI	
First presentation											
VF	0	47	13.87	8.83	1.05	.11	.91	.22	.88	.86	.90
MF	1	43	15.42	8.58	.70	.16	-.05	.32	.87	.84	.89
SF	1	46	16.71	9.28	.72	.21	.22	.42	.89	.86	.91
LF	2	46	18.81	9.75	.49	.23	-.43	.46	.87	.83	.91
Second presentations											
VF	0	46	12.95	8.99	1.12	.11	1.14	.22	.90	.89	.91
MF	0	43	14.36	8.66	.82	.16	.38	.32	.88	.85	.90
SF	0	46	15.79	9.51	.87	.21	.52	.42	.89	.87	.92
LF	0	43	17.23	10.08	.63	.23	-.16	.46	.88	.84	.91
One item per screen											
VF	0	47	13.33	8.99	1.11	.11	1.17	.22	.90	.88	.91
MF	0	43	14.85	8.71	.77	.16	.12	.32	.88	.85	.90
SF	1	46	16.21	9.23	.86	.21	.65	.42	.89	.85	.91
LF	0	46	18.01	9.95	.56	.23	-.28	.46	.88	.84	.91
Multiple items per screen											
VF	0	45	13.48	8.85	1.05	.11	.85	.22	.88	.87	.90
MF	0	43	14.94	8.56	.733	.16	.18	.32	.87	.84	.89
SF	0	44	16.29	9.57	.730	.21	.10	.42	.90	.87	.92
LF	2	43	18.02	9.94	.55	.23	-.35	.46	.88	.84	.91

Note. $n_{\text{VF}} = 475$; $n_{\text{MF}} = 223$; $n_{\text{SF}} = 134$; $n_{\text{LF}} = 108$.

VF = very fluent, English is my first language; MF = more fluent in English than my first language; SF = same fluency in English as my first language; LF = less fluent in English than first my language.

Table 11. Descriptive Statistics of Subscale Scores for Males and Females on Each Assessment

	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
First Presentation								
Somatic Symptoms								
Male	0	18	6.49	3.23	.56	.14	.38	.27
Female	0	19	6.57	3.38	.61	.10	.38	.20
Depressed Affect								
Male	0	19	3.28	3.40	1.45	.14	2.34	.27
Female	0	20	4.48	4.16	1.06	.10	.59	.20
Positive Affect								
Male	0	12	3.78	2.91	.62	.14	-.25	.27
Female	0	12	3.75	2.82	.48	.10	-.56	.20
Interpersonal Problems								
Male	0	6	.85	1.12	1.58	.14	2.93	.27
Female	0	6	.84	1.13	1.44	.10	1.78	.20
Second Presentation								
Somatic Symptoms								
Male	0	18	6.08	3.39	.65	.14	.62	.27
Female	0	19	6.12	3.49	.52	.10	.28	.20
Depressed Affect								
Male	0	16	2.84	3.44	1.66	.14	2.82	.27
Female	0	20	4.12	4.24	1.28	.10	1.23	.20
Positive Affect								
Male	0	12	3.59	2.99	.59	.14	-.38	.27
Female	0	12	3.66	2.89	.46	.10	-.62	.20
Interpersonal Problems								
Male	0	6	.76	1.15	1.80	.14	3.65	.27
Female	0	6	.76	1.14	1.64	.10	2.46	.20

Note. n_{male} = 324; n_{female} = 616.

Table 12. Descriptive Statistics of Subscale CES-D Scores for Males and Females for each Screen Format

	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
One item per screen								
Somatic Symptoms								
Male	0	18	6.37	3.32	.54	.14	0.46	.27
Female	0	19	6.39	3.40	.55	.10	0.33	.20
Depressed Affect								
Male	0	19	3.10	3.51	1.62	.14	2.90	.27
Female	0	20	4.30	4.22	1.18	.10	0.91	.20
Positive Affect								
Male	0	12	3.57	2.91	.58	.14	-0.38	.27
Female	0	12	3.59	2.85	.51	.10	-0.51	.20
Interpersonal Problems								
Male	0	6	.80	1.13	1.80	.14	4.05	.27
Female	0	6	.78	1.12	1.52	.10	1.94	.20
Multiple items per screen								
Somatic Symptoms								
Male	0	18	6.20	3.32	.65	.14	0.55	.27
Female	0	19	6.30	3.49	.61	.10	0.32	.20
Depressed Affect								
Male	0	16	3.03	3.34	1.45	.14	2.03	.27
Female	0	20	4.30	4.19	1.16	.10	0.88	.20
Positive Affect								
Male	0	12	3.79	2.99	.62	.14	-0.27	.27
Female	0	11	3.82	2.86	.43	.10	-0.66	.20
Interpersonal Problems								
Male	0	6	.81	1.15	1.58	.14	2.55	.27
Female	0	6	.82	1.15	1.55	.10	2.23	.20

Note. $n_{\text{male}} = 324$; $n_{\text{female}} = 616$.

Table 13. **Descriptive Statistics of Subscale Scores for Participants with Different Levels of Self-rated English Fluency on each Assessment**

	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
First Presentation								
Somatic Symptoms								
VF	0	19	6.34	3.43	.63	.11	.49	.22
MF	0	16	6.54	3.23	.55	.16	-.02	.32
SF	0	18	6.70	3.27	.82	.21	.98	.42
LF	0	17	7.21	3.11	.40	.23	.24	.46
Depressed Affect								
VF	0	18	3.63	3.86	1.40	.11	1.62	.22
MF	0	16	3.88	3.67	1.11	.16	.68	.32
SF	0	20	4.71	4.04	1.12	.21	1.55	.42
LF	0	19	5.57	4.42	.75	.23	-.07	.46
Positive Affect								
VF	0	12	3.07	2.55	.69	.11	-.06	.22
MF	0	12	4.21	2.77	.37	.16	-.52	.32
SF	0	11	4.41	2.91	.18	.21	-.88	.42
LF	0	12	5.09	3.34	.21	.23	-1.00	.46
Interpersonal Problems								
VF	0	6	.84	1.11	1.47	.11	2.02	.22
MF	0	6	.79	1.11	1.61	.16	2.74	.32
SF	0	5	.90	1.13	1.36	.21	1.41	.42
LF	0	6	.94	1.23	1.49	.23	2.51	.46
Second Presentation								
Somatic Symptoms								
VF	0	19	5.93	3.53	.60	.11	.28	.22
MF	0	17	6.00	3.37	.76	.16	.78	.32
SF	0	18	6.51	3.40	.61	.21	.65	.42
LF	0	17	6.60	3.32	.42	.23	.27	.46

	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Depressed Affect								
VF	0	20	3.29	3.87	1.56	.11	2.20	.22
MF	0	17	3.50	3.72	1.40	.16	1.74	.32
SF	0	20	4.14	4.24	1.44	.21	2.18	.42
LF	0	18	5.17	4.66	.87	.23	.06	.46
Positive Affect								
VF	0	12	2.99	2.69	.78	.11	.08	.22
MF	0	12	4.11	2.89	.24	.16	-.67	.32
SF	0	11	4.29	2.90	.17	.21	-.76	.42
LF	0	12	4.68	3.35	.23	.23	-1.1	.46
Interpersonal Problems								
VF	0	6	.74	1.14	1.88	.11	4.00	.22
MF	0	6	.75	1.11	1.67	.16	2.91	.32
SF	0	4	.84	1.21	1.24	.21	.45	.42
LF	0	5	.75	1.14	1.56	.23	1.94	.46

Note. n_{VF} = 475; n_{MF} = 223; n_{SF} = 134; n_{LF} = 108.

VF = very fluent, English is my first language; MF = more fluent in English than my first language; SF = same fluency in English as my first language; LF = less fluent in English than first my language.

Table 14. Descriptive Statistics of Subscale Scores for Participants with Different Levels of Self-rated English Fluency for each Screen Format

	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
One item per screen								
Somatic Symptoms								
VF	0	19	6.17	3.45	.55	.11	.33	.22
MF	0	17	6.34	3.30	.58	.16	.24	.32
SF	0	18	6.70	3.27	.81	.21	1.07	.42
LF	0	17	6.97	3.24	.36	.23	.28	.46
Depressed Affect								
VF	0	20	3.48	3.94	1.51	.11	1.98	.22
MF	0	17	3.71	3.76	1.29	.16	1.31	.32
SF	0	20	4.40	4.08	1.31	.21	2.05	.42
LF	0	19	5.40	4.52	.77	.23	.00	.46
Positive Affect								
VF	0	12	2.90	2.63	.80	.11	.14	.22
MF	0	12	4.04	2.81	.29	.16	-.66	.32
SF	0	11	4.28	2.84	.18	.21	-.80	.42
LF	0	12	4.78	3.26	.25	.23	-.90	.46
Interpersonal Problems								
VF	0	6	.77	1.11	1.76	.11	3.57	.22
MF	0	6	.77	1.11	1.64	.16	2.86	.32
SF	0	4	.82	1.16	1.31	.21	.74	.42
LF	0	5	.84	1.15	1.42	.23	1.67	.46
Multiple items per screen								
Somatic Symptoms								
VF	0	19	6.09	3.52	.66	.11	.42	.22
MF	0	17	6.20	3.32	.72	.16	.50	.32
SF	0	18	6.52	3.40	.62	.21	.58	.42
LF	0	17	6.84	3.22	.42	.23	.22	.46

	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Depressed Affect								
VF	0	18	3.44	3.79	1.43	.11	1.76	.22
MF	0	16	3.67	3.64	1.20	.16	1.02	.32
SF	0	20	4.45	4.22	1.24	.21	1.61	.42
LF	0	18	5.34	4.58	0.84	.23	-.02	.46
Positive Affect								
VF	0	12	3.15	2.61	0.68	.11	-.08	.22
MF	0	12	4.28	2.85	0.31	.16	-.53	.32
SF	0	11	4.42	2.97	0.16	.21	-.84	.42
LF	0	12	4.99	3.44	0.18	.23	-1.17	.46
Interpersonal Problems								
VF	0	6	.80	1.14	1.60	.11	2.51	.22
MF	0	6	.77	1.11	1.64	.16	2.79	.32
SF	0	5	.92	1.18	1.29	.21	.99	.42
LF	0	6	.85	1.23	1.61	.23	2.76	.46

Note. $n_{\text{VF}} = 475$; $n_{\text{MF}} = 223$; $n_{\text{SF}} = 134$; $n_{\text{LF}} = 108$.

VF = very fluent, English is my first language; MF = more fluent in English than my first language; SF = same fluency in English as my first language; LF = less fluent in English than first my language.

Table 15. Descriptive Statistics of Item Responses for Participants with Different Levels of Self-rated English Fluency on the First Assessment

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 1									
VF	474	0	3	.76	.78	.74	.11	-.10	.22
MF	223	0	3	.76	.74	.68	.16	.05	.32
SF	134	0	3	.80	.78	.75	.21	.13	.42
LF	108	0	3	1.04	.99	.58	.23	-.71	.46
Item 2									
VF	474	0	3	.54	.78	1.38	.11	1.27	.22
MF	221	0	3	.59	.78	1.10	.16	.30	.33
SF	134	0	3	.55	.73	1.28	.21	1.35	.42
LF	108	0	3	.53	.79	1.41	.23	1.22	.46
Item 3									
VF	468	0	3	.63	.83	1.17	.11	.51	.23
MF	220	0	3	.60	.80	1.23	.16	.80	.33
SF	134	0	3	.76	.89	1.00	.21	.17	.42
LF	108	0	3	.81	.82	.67	.23	-.36	.46
Item 4									
VF	472	0	3	.75	.86	.86	.11	-.24	.22
MF	220	0	3	1.12	.97	.36	.16	-.94	.33
SF	131	0	3	1.22	1.00	.43	.21	-.85	.42
LF	108	0	3	1.29	1.08	.23	.23	-1.23	.46
Item 5									
VF	475	0	3	1.51	.87	.03	.11	-.68	.22
MF	221	0	3	1.53	.86	.04	.16	-.64	.33
SF	134	0	3	1.35	.93	.16	.21	-.81	.42
LF	107	0	3	1.52	.89	-.11	.23	-.71	.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 6									
VF	474	0	3	.52	.80	1.52	.11	1.61	.22
MF	223	0	3	.70	.83	.98	.16	.22	.32
SF	134	0	3	.81	.84	.91	.21	.30	.42
LF	108	0	3	1.18	.95	.45	.23	-.65	.46
Item 7									
VF	474	0	3	1.11	.90	.45	.11	-.56	.22
MF	223	0	3	1.29	.89	.24	.16	-.68	.32
SF	132	0	3	1.45	.86	.07	.21	-.61	.42
LF	108	0	3	1.47	.97	.24	.23	-.93	.46
Item 8									
VF	471	0	3	.91	.84	.52	.11	-.60	.23
MF	222	0	3	1.27	.90	.28	.16	-.67	.33
SF	134	0	3	1.11	.91	.31	.21	-.84	.42
LF	107	0	3	1.31	1.01	.18	.23	-1.08	.46
Item 9									
VF	472	0	3	.23	.52	2.45	.11	5.94	.22
MF	223	0	3	.30	.56	2.06	.16	4.87	.32
SF	134	0	3	.46	.74	1.59	.21	1.91	.42
LF	107	0	3	.60	.87	1.24	.23	.45	.46
Item 10									
VF	474	0	3	.54	.74	1.18	.11	.60	.22
MF	223	0	3	.49	.76	1.60	.16	2.03	.32
SF	134	0	3	.57	.76	1.12	.21	.39	.42
LF	108	0	3	.55	.85	1.50	.23	1.39	.46
Item 11									
VF	475	0	3	.97	.95	.66	.11	-.56	.22
MF	223	0	3	.87	.91	.76	.16	-.35	.32
SF	134	0	3	.89	.90	.67	.21	-.49	.42
LF	107	0	3	.95	.90	.56	.23	-.61	.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 12									
VF	469	0	3	.69	.75	.81	.11	.01	.23
MF	222	0	3	.92	.82	.60	.16	-.21	.33
SF	133	0	3	1.03	.81	.30	.21	-.61	.42
LF	107	0	3	1.28	1.00	.11	.23	-1.12	.46
Item 13									
VF	470	0	3	.70	.77	.88	.11	.16	.23
MF	220	0	3	.72	.75	.70	.16	-.26	.33
SF	133	0	3	.89	.81	.73	.21	.14	.42
LF	108	0	3	1.03	.98	.49	.23	-.89	.46
Item 14									
VF	472	0	3	.76	.88	.96	.11	.07	.22
MF	223	0	3	.80	.89	.90	.16	-.01	.32
SF	133	0	3	.86	.88	.83	.21	-.03	.42
LF	107	0	3	1.11	1.02	.47	.23	-.93	.46
Item 15									
VF	474	0	3	.42	.66	1.53	.11	1.94	.22
MF	223	0	3	.34	.62	1.87	.16	3.17	.32
SF	134	0	3	.43	.68	1.58	.21	2.26	.42
LF	107	0	3	.46	.72	1.56	.23	1.97	.46
Item 16									
VF	473	0	3	.70	.79	.85	.11	-.13	.22
MF	222	0	3	.91	.89	.58	.16	-.66	.33
SF	132	0	3	1.05	.91	.39	.21	-.80	.42
LF	108	0	3	1.20	1.01	.24	.23	-1.11	.46
Item 17									
VF	474	0	3	.27	.61	2.43	.11	5.76	.22
MF	222	0	3	.28	.63	2.47	.16	6.10	.33
SF	134	0	3	.36	.68	2.09	.21	4.30	.42
LF	108	0	3	.45	.79	1.78	.23	2.47	.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 18									
VF	472	0	3	.68	.80	1.05	.11	.55	.22
MF	221	0	3	.70	.79	.82	.16	-.21	.33
SF	133	0	3	.90	.83	.66	.21	-.13	.42
LF	108	0	3	.88	.90	.86	.23	.02	.46
Item 19									
VF	473	0	3	.42	.68	1.63	.11	2.31	.22
MF	221	0	3	.44	.70	1.59	.16	2.22	.33
SF	134	0	3	.46	.66	1.27	.21	1.12	.42
LF	108	0	3	.48	.70	1.46	.23	1.90	.46
Item 20									
VF	473	0	3	.75	.81	.84	.11	.02	.22
MF	223	0	3	.79	.89	.96	.16	.12	.32
SF	134	0	3	.78	.84	1.04	.21	.67	.42
LF	108	0	3	.68	.88	1.19	.23	.55	.46

Note. VF = very fluent, English is my first language; MF = more fluent in English than my first language; SF = same fluency in English as my first language; LF = less fluent in English than first my language.

Table 16. **Descriptive Statistics of Item Responses for Participants with Different Levels of Self-rated English Fluency on the Second Assessment**

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 1									
VF	474	0	3	.63	.72	.88	.11	.22	.22
MF	223	0	3	.68	.69	.69	.16	.02	.32
SF	134	0	3	.75	.77	.76	.21	.03	.42
LF	107	0	3	.87	.91	.80	.23	-.23	.46
Item 2									
VF	467	0	3	.51	.75	1.46	.11	1.59	.23
MF	222	0	3	.53	.73	1.20	.16	.71	.33
SF	133	0	3	.59	.74	1.16	.21	1.02	.42
LF	107	0	3	.44	.69	1.63	.23	2.54	.46
Item 3									
VF	468	0	3	.51	.76	1.44	.11	1.46	.23
MF	221	0	3	.50	.70	1.29	.16	1.18	.33
SF	132	0	3	.65	.79	1.09	.21	.62	.42
LF	104	0	3	.73	.75	.63	.24	-.49	.47
Item 4									
VF	473	0	3	.78	.89	.85	.11	-.28	.22
MF	222	0	3	1.09	.95	.33	.16	-.99	.33
SF	134	0	3	1.16	.91	.36	.21	-.68	.42
LF	107	0	3	1.21	.93	.22	.23	-.87	.46
Item 5									
VF	471	0	3	1.36	.86	.21	.11	-.56	.23
MF	222	0	3	1.34	.86	.37	.16	-.43	.33
SF	133	0	3	1.31	.93	.15	.21	-.86	.42
LF	108	0	3	1.41	.95	.10	.23	-.88	.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 6									
VF	473	0	3	.49	.79	1.63	.11	1.98	.22
MF	222	0	3	.61	.84	1.20	.16	.50	.33
SF	132	0	3	.68	.85	1.19	.21	.81	.42
LF	108	0	3	.98	1.01	.70	.23	-.66	.46
Item 7									
VF	475	0	3	1.04	.91	.58	.11	-.44	.22
MF	223	0	3	1.19	.90	.33	.16	-.65	.32
SF	134	0	3	1.35	.98	.23	.21	-.91	.42
LF	108	0	3	1.37	.99	.25	.23	-.94	.46
Item 8									
VF	473	0	3	.88	.86	.65	.11	-.39	.22
MF	221	0	3	1.23	.88	.17	.16	-.76	.33
SF	134	0	3	1.09	.95	.31	.21	-1.00	.42
LF	107	0	3	1.18	1.02	.35	.23	-1.02	.46
Item 9									
VF	472	0	3	.19	.47	2.61	.11	6.89	.22
MF	221	0	3	.30	.58	2.06	.16	4.42	.33
SF	133	0	3	.39	.68	1.63	.21	1.74	.42
LF	108	0	3	.54	.88	1.56	.23	1.42	.46
Item 10									
VF	473	0	3	.47	.74	1.55	.11	1.86	.22
MF	221	0	3	.40	.68	1.77	.16	2.94	.33
SF	134	0	3	.47	.73	1.44	.21	1.30	.42
LF	108	0	3	.53	.86	1.72	.23	2.22	.46
Item 11									
VF	471	0	3	.96	.93	.65	.11	-.51	.23
MF	221	0	3	.84	.91	.84	.16	-.20	.33
SF	133	0	3	.90	.94	.76	.21	-.37	.42
LF	108	0	3	.92	.96	.69	.23	-.60	.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 12									
VF	473	0	3	.67	.79	.96	.11	.21	.22
MF	219	0	3	.90	.84	.67	.16	-.20	.33
SF	133	0	3	1.04	.83	.33	.21	-.62	.42
LF	107	0	3	1.22	.96	.18	.23	-1.02	.46
Item 13									
VF	474	0	3	.65	.72	.95	.11	.58	.22
MF	223	0	3	.66	.73	.76	.16	-.27	.32
SF	133	0	3	.77	.72	.49	.21	-.52	.42
LF	108	0	3	.86	.88	.78	.23	-.15	.46
Item 14									
VF	471	0	3	.69	.83	1.03	.11	.34	.23
MF	223	0	3	.72	.88	1.09	.16	.38	.32
SF	134	0	3	.81	.86	.95	.21	.34	.42
LF	106	0	3	1.00	1.02	.65	.24	-.75	.47
Item 15									
VF	474	0	3	.38	.65	1.71	.11	2.52	.22
MF	223	0	3	.37	.62	1.71	.16	2.70	.32
SF	133	0	3	.38	.63	1.65	.21	2.28	.42
LF	107	0	3	.38	.71	2.04	.23	4.01	.46
Item 16									
VF	472	0	3	.65	.79	.98	.11	.14	.22
MF	222	0	3	.88	.84	.55	.16	-.61	.33
SF	133	0	3	.02	.84	.43	.21	-.51	.42
LF	108	0	3	1.10	.99	.32	.23	-1.12	.46
Item 17									
VF	475	0	3	.29	.62	2.35	.11	5.37	.22
MF	223	0	3	.29	.61	2.33	.16	5.40	.32
SF	134	0	3	.33	.69	2.37	.21	5.49	.42
LF	107	0	3	.49	.81	1.65	.23	1.96	.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 18									
VF	473	0	3	.65	.80	1.07	.11	.46	.22
MF	220	0	3	.66	.80	1.12	.16	.77	.33
SF	133	0	3	.79	.90	1.01	.21	.25	.42
LF	108	0	3	.89	.91	.83	.23	-.08	.46
Item 19									
VF	473	0	3	.36	.65	1.93	.11	3.77	.22
MF	222	0	3	.39	.63	1.61	.16	2.37	.33
SF	133	0	3	.47	.71	1.30	.21	.71	.42
LF	108	0	2	.37	.62	1.47	.23	1.03	.46
Item 20									
VF	472	0	3	.78	.81	.79	.11	-.08	.22
MF	223	0	3	.76	.88	1.02	.16	.28	.32
SF	133	0	3	.83	.92	.83	.21	-.28	.42
LF	108	0	3	.72	.89	1.06	.23	.22	.46

Note. VF = very fluent, English is my first language; MF = more fluent in English than my first language; SF = same fluency in English as my first language; LF = less fluent in English than first my language.

Table 17. Descriptive Statistics of Item Responses for Participants with Different Levels of Self-rated English Fluency on the One Item per Screen Format

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 1									
VF	474	0	3	.70	.75	.76	.11	-.17	.22
MF	223	0	3	.73	.70	.59	.16	-.17	.32
SF	134	0	3	.84	.79	.68	.21	-.02	.42
LF	107	0	3	.98	.99	.69	.23	-.59	.46
Item 2									
VF	472	0	3	.51	.75	1.46	.11	1.64	.22
MF	223	0	3	.56	.74	1.18	.16	.75	.32
SF	133	0	3	.54	.71	1.32	.21	1.68	.42
LF	108	0	3	.49	.73	1.43	.23	1.51	.46
Item 3									
VF	469	0	3	.57	.80	1.31	.11	1.01	.23
MF	221	0	3	.52	.73	1.31	.16	1.17	.33
SF	133	0	3	.69	.82	1.05	.21	.49	.42
LF	104	0	3	.75	.77	.72	.24	-.14	.47
Item 4									
VF	474	0	3	.72	.84	.87	.11	-.24	.22
MF	220	0	3	1.10	.96	.37	.16	-.91	.33
SF	133	0	3	1.21	.97	.43	.21	-.75	.42
LF	107	0	3	1.22	.97	.25	.23	-.97	.46
Item 5									
VF	472	0	3	1.46	.86	.14	.11	-.62	.22
MF	222	0	3	1.45	.84	.24	.16	-.52	.33
SF	133	0	3	1.37	.89	.05	.21	-.75	.42
LF	108	0	3	1.46	.90	-.04	.23	-.75	.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 6									
VF	473	0	3	.50	.80	1.58	.11	1.78	.22
MF	222	0	3	.66	.85	1.13	.16	.44	.33
SF	133	0	3	.74	.86	1.03	.21	.40	.42
LF	108	0	3	1.11	.98	.50	.23	-.74	.46
Item 7									
VF	475	0	3	1.10	.92	.50	.11	-.56	.22
MF	223	0	3	1.27	.89	.26	.16	-.65	.32
SF	134	0	3	1.42	.93	.16	.21	-.80	.42
LF	108	0	3	1.43	.97	.25	.23	-.89	.46
Item 8									
VF	473	0	3	.86	.85	.60	.11	-.55	.22
MF	222	0	3	1.23	.89	.20	.16	-.75	.33
SF	134	0	3	1.08	.95	.37	.21	-.92	.42
LF	107	0	3	1.23	1.01	.29	.23	-1.04	.46
Item 9									
VF	471	0	3	.20	.49	2.62	.11	7.31	.23
MF	223	0	3	.31	.59	1.99	.16	4.10	.32
SF	133	0	3	.44	.72	1.58	.21	1.74	.42
LF	107	0	3	.53	.86	1.52	.23	1.32	.46
Item 10									
VF	472	0	3	.50	.75	1.38	.11	1.17	.22
MF	223	0	3	.44	.74	1.80	.16	2.83	.32
SF	134	0	3	.49	.71	1.23	.21	.57	.42
LF	108	0	3	.53	.84	1.57	.23	1.67	.46
Item 11									
VF	471	0	3	.97	.93	.67	.11	-.45	.23
MF	221	0	3	.86	.90	.78	.16	-.30	.33
SF	133	0	3	.92	.93	.68	.21	-.50	.42
LF	107	0	3	.94	.92	.71	.23	-.34	.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 12									
VF	473	0	3	.67	.76	.91	.11	.18	.22
MF	222	0	3	.86	.80	.68	.16	.04	.33
SF	133	0	3	1.02	.81	.39	.21	-.43	.42
LF	107	0	3	1.22	.95	.16	.23	-1.02	.46
Item 13									
VF	473	0	3	.67	.75	.92	.11	.35	.22
MF	222	0	3	.68	.73	.72	.16	-.30	.33
SF	134	0	3	.84	.75	.61	.21	.04	.42
LF	108	0	3	0.94	0.92	0.63	.23	-0.50	.46
Item 14									
VF	473	0	3	.75	.87	.99	.11	.15	.22
MF	223	0	3	.79	.92	.96	.16	-.02	.32
SF	134	0	3	.82	.86	.86	.21	.11	.42
LF	107	0	3	1.07	1.03	.49	.23	-.96	.46
Item 15									
VF	473	0	3	.40	.64	1.55	.11	1.97	.22
MF	223	0	3	.35	.61	1.67	.16	2.19	.32
SF	134	0	2	.36	.61	1.49	.21	1.14	.42
LF	106	0	3	.44	.69	1.62	.24	2.49	.47
Item 16									
VF	473	0	3	.65	.78	.94	.11	.02	.22
MF	222	0	3	.83	.83	.62	.16	-.51	.33
SF	133	0	3	.99	.82	.37	.21	-.62	.42
LF	108	0	3	1.13	.98	.29	.23	-1.06	.46
Item 17									
VF	474	0	3	.28	.61	2.37	.11	5.46	.22
MF	222	0	3	.28	.62	2.47	.16	6.28	.33
SF	134	0	3	.36	.69	2.08	.21	4.09	.42
LF	107	0	3	.49	.83	1.72	.23	2.16	.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 18									
VF	473	0	3	.69	.83	1.07	.11	.45	.22
MF	221	0	3	.69	.78	.95	.16	.38	.33
SF	133	0	3	.85	.87	.80	.21	-.07	.42
LF	108	0	3	.91	.88	.77	.23	-.07	.46
Item 19									
VF	474	0	3	.38	.65	1.82	.11	3.21	.22
MF	222	0	3	.42	.65	1.59	.16	2.44	.33
SF	133	0	2	.47	.68	1.15	.21	.05	.42
LF	108	0	2	.40	.61	1.28	.23	.59	.46
Item 20									
VF	473	0	3	.76	.82	.84	.11	-.02	.22
MF	223	0	3	.81	.88	.91	.16	.10	.32
SF	133	0	3	.78	.85	.89	.21	.12	.42
LF	108	0	3	.72	.91	1.12	.23	.39	.46

Note. VF = very fluent, English is my first language; MF = more fluent in English than my first language; SF = same fluency in English as my first language; LF = less fluent in English than first my language.

Table 18. **Descriptive Statistics of Item Responses for Participants with Different Levels of Self-rated English Fluency on the Multiple Items per Screen Format**

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 1									
VF	474	0	3	.69	.75	.88	.11	.31	.22
MF	223	0	3	.71	.73	.78	.16	.27	.32
SF	134	0	3	.72	.76	.84	.21	.23	.42
LF	108	0	3	.93	.91	.67	.23	-.44	.46
Item 2									
VF	469	0	3	.54	.77	1.37	.11	1.22	.23
MF	220	0	3	.56	.77	1.13	.16	.27	.33
SF	134	0	3	.60	.76	1.12	.21	.75	.42
LF	107	0	3	.48	.76	1.61	.23	2.10	.46
Item 3									
VF	467	0	3	.57	.80	1.29	.11	.86	.23
MF	220	0	3	.58	.77	1.24	.16	.95	.33
SF	133	0	3	.72	.87	1.06	.21	.33	.42
LF	108	0	3	.80	.81	.61	.23	-.56	.46
Item 4									
VF	471	0	3	.81	.91	.83	.11	-.33	.23
MF	222	0	3	1.11	.97	.32	.16	-1.01	.33
SF	132	0	3	1.17	.94	.38	.21	-.75	.42
LF	108	0	3	1.28	1.04	.23	.23	-1.13	.46
Item 5									
VF	474	0	3	1.41	.87	.11	.11	-.66	.22
MF	221	0	3	1.43	.88	.17	.16	-.67	.33
SF	134	0	3	1.29	.96	.25	.21	-.88	.42
LF	107	0	3	1.47	.94	.03	0.23	-0.88	0.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 6									
VF	474	0	3	.50	.79	1.57	.11	1.80	.22
MF	223	0	3	.66	.83	1.05	.16	.22	.32
SF	133	0	3	.74	.84	1.06	.21	.62	.42
LF	108	0	3	1.05	.99	.61	.23	-.65	.46
Item 7									
VF	474	0	3	1.05	.89	.53	.11	-.46	.22
MF	223	0	3	1.21	.90	.31	.16	-.68	.32
SF	132	0	3	1.39	.91	.13	.21	-.77	.42
LF	108	0	3	1.42	1.00	.24	.23	-.98	.46
Item 8									
VF	471	0	3	.92	.85	.58	.11	-.44	.23
MF	221	0	3	1.28	.90	.26	.16	-.67	.33
SF	134	0	3	1.12	.91	.25	.21	-.92	.42
LF	107	0	3	1.25	1.02	.24	.23	-1.09	.46
Item 9									
VF	473	0	3	.21	.51	2.46	.11	5.67	.22
MF	221	0	3	.29	.55	2.14	.16	5.27	.33
SF	134	0	3	.42	.71	1.65	.21	2.07	.42
LF	108	0	3	.60	.89	1.29	.23	0.59	.46
Item 10									
VF	475	0	3	.52	.73	1.33	.11	1.20	.22
MF	221	0	3	.45	.70	1.56	.16	2.04	.33
SF	134	0	3	.54	.78	1.29	.21	.83	.42
LF	108	0	3	.55	.87	1.64	.23	1.91	.46
Item 11									
VF	475	0	3	.96	.95	.64	.11	-.62	.22
MF	223	0	3	.85	.92	.82	.16	-.26	.32
SF	134	0	3	.87	.90	.75	.21	-.33	.42
LF	108	0	3	.93	.94	.56	.23	-.86	.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 12									
VF	469	0	3	.70	.78	.87	.11	.06	.23
MF	219	0	3	.95	.86	.57	.16	-.42	.33
SF	133	0	3	1.05	.82	.25	.21	-.78	.42
LF	107	0	3	1.29	1.01	.12	.23	-1.13	.46
Item 13									
VF	471	0	3	.67	.75	.92	.11	.38	.23
MF	221	0	3	.71	.76	.74	.16	-.25	.33
SF	132	0	3	.83	.80	.69	.21	-.06	.42
LF	108	0	3	.94	.96	0.64	.23	-.65	.46
LF	108	0	3	.86	.88	.78	.23	-.15	.46
Item 14									
VF	470	0	3	.70	.83	1.00	.11	.24	.23
MF	223	0	3	.74	.85	1.02	.16	.35	.32
SF	133	0	3	.85	.88	.90	.21	.17	.42
LF	106	0	3	1.04	1.02	.63	.24	-.74	.47
Item 15									
VF	475	0	3	.40	.66	1.67	.11	2.42	.22
MF	223	0	3	.36	.63	1.88	.16	3.53	.32
SF	133	0	3	.45	.70	1.65	.21	2.63	.42
LF	108	0	3	.40	.74	1.94	.23	3.27	.46
Item 16									
VF	472	0	3	.70	.80	.89	.11	-.03	.22
MF	222	0	3	.96	.90	.50	.16	-.76	.33
SF	132	0	3	1.08	.93	.40	.21	-.79	.42
LF	108	0	3	1.18	1.03	.27	.23	-1.17	.46

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 17									
VF	475	0	3	.28	.62	2.41	.11	5.65	.22
MF	223	0	3	.29	.62	2.33	.16	5.26	.32
SF	134	0	3	.33	.68	2.39	.21	5.79	.42
LF	108	0	3	.45	.77	1.68	.23	2.16	0.46
Item 18									
VF	472	0	3	.64	.76	1.03	.11	.49	.22
MF	220	0	3	.67	.81	.99	.16	.19	.33
SF	133	0	3	.84	.87	.88	.21	.15	.42
LF	108	0	3	.86	.93	.92	.23	.00	.46
Item 19									
VF	472	0	3	.41	.67	1.72	.11	2.74	.22
MF	221	0	3	.41	.67	1.64	.16	2.31	.33
SF	134	0	3	.47	.69	1.43	.21	1.73	.42
LF	108	0	3	.45	.72	1.57	.23	2.02	.46
Item 20									
VF	472	0	3	.77	.80	.79	.11	-.05	.22
MF	223	0	3	.74	.89	1.07	.16	.33	.32
SF	134	0	3	.83	.91	.95	.21	.09	.42
LF	108	0	3	.68	.87	1.12	.23	.34	.46

Note. VF = very fluent, English is my first language; MF = more fluent in English than my first language; SF = same fluency in English as my first language; LF = less fluent in English than first my language.

Table 19. Descriptive Statistics of Item Responses for Males and Females on the First Assessment

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 1									
Male	324	0	3	.71	.80	1.06	.14	.74	.27
Female	615	0	3	.85	.80	.61	.10	-.30	.20
Item 2									
Male	323	0	3	.48	.73	1.47	.14	1.54	.27
Female	614	0	3	.59	.79	1.22	.10	.77	.20
Item 3									
Male	317	0	3	.62	.82	1.19	.14	.66	.27
Female	613	0	3	.69	.84	1.04	.10	.24	.20
Item 4									
Male	321	0	3	.93	.98	.73	.14	-.58	.27
Female	610	0	3	.99	.95	.55	.10	-.75	.20
Item 5									
Male	323	0	3	1.46	.86	.03	.14	-.65	.27
Female	614	0	3	1.51	.89	.02	.10	-.74	.20
Item 6									
Male	324	0	3	.65	.87	1.26	.14	.78	.27
Female	615	0	3	.69	.85	1.05	.10	.29	.20
Item 7									
Male	321	0	3	1.30	.90	.20	.14	-.72	.27
Female	616	0	3	1.22	.92	.39	.10	-.65	.20
Item 8									
Male	322	0	3	1.08	.90	.44	.14	-.64	.27
Female	612	0	3	1.07	.91	.40	.10	-.75	.20
Item 9									
Male	324	0	3	.35	.66	1.99	.14	3.58	.27
Female	612	0	3	.30	.60	2.10	.10	4.24	.20
Item 10									
Male	324	0	3	.22	.49	2.55	.14	7.65	.27
Female	615	0	3	.70	.82	.93	.10	.03	.20

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 11									
Male	324	0	3	.94	.96	.73	.14	-.50	.27
Female	615	0	3	.93	.91	.64	.10	-.54	.20
Item 12									
Male	321	0	3	.88	.82	.64	.14	-.20	.27
Female	610	0	3	.85	.84	.63	.10	-.43	.20
Item 13									
Male	322	0	3	.81	.81	.78	.14	.10	.27
Female	609	0	3	.75	.81	.81	.10	-.11	.20
Item 14									
Male	322	0	3	.81	.90	.91	.14	-.04	.27
Female	613	0	3	.83	.90	.85	.10	-.19	.20
Item 15									
Male	323	0	3	.40	.65	1.72	.14	3.00	.27
Female	615	0	3	.41	.66	1.56	.10	1.85	.20
Item 16									
Male	323	0	3	.89	.89	.59	.14	-.68	.27
Female	612	0	3	.84	.87	.71	.10	-.44	.20
Item 17									
Male	323	0	2	.09	.35	4.13	.14	17.35	.27
Female	615	0	3	.42	.73	1.82	.10	2.81	.20
Item 18									
Male	323	0	3	.55	.74	1.32	.14	1.34	.27
Female	611	0	3	.84	.84	.75	0.10	-0.11	0.20
Item 19									
Male	321	0	3	.45	.69	1.51	.14	1.87	.27
Female	615	0	3	.43	.67	1.57	.10	2.15	.20
Item 20									
Male	323	0	3	.81	.87	.84	.14	-.11	.27
Female	615	0	3	.73	.83	1.01	.10	.41	.20

Table 20. Descriptive Statistics of Item Responses for Males and Females on the Second Assessment

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 1									
Male	323	0	3	.58	.72	1.19	.14	1.15	.27
Female	615	0	3	.75	.75	.71	.10	-.06	.20
Item 2									
Male	320	0	3	.44	.69	1.61	.14	2.40	.27
Female	609	0	3	.56	.76	1.25	.10	.94	.20
Item 3									
Male	319	0	3	.47	.70	1.39	.14	1.26	.27
Female	606	0	3	.60	.77	1.18	.10	.78	.20
Item 4									
Male	324	0	3	.89	.92	.69	.14	-.54	.27
Female	612	0	3	.99	.93	.49	.10	-.81	.20
Item 5									
Male	322	0	3	1.36	.88	.22	.14	-.64	.27
Female	612	0	3	1.35	.87	.22	.10	-.62	.20
Item 6									
Male	323	0	3	.54	.82	1.44	.14	1.23	.27
Female	612	0	3	.63	.87	1.27	.10	.75	.20
Item 7									
Male	324	0	3	1.22	.93	.32	.14	-.77	.27
Female	616	0	3	1.13	.93	.50	.10	-.59	.20
Item 8									
Male	321	0	3	1.05	.93	.47	.14	-.72	.27
Female	614	0	3	1.02	.90	.44	.10	-.74	.20
Item 9									
Male	324	0	3	.33	.67	2.23	.14	4.80	.27
Female	610	0	3	.26	.56	2.20	.10	4.50	.20
Item 10									
Male	323	0	3	.15	.43	3.35	.14	13.72	.27
Female	613	0	3	.62	.81	1.19	.10	.74	.20

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 11									
Male	322	0	3	.93	.95	.73	.14	-.46	.27
Female	611	0	3	.91	.92	.70	.10	-.45	.20
Item 12									
Male	322	0	3	.84	.85	.70	.14	-.28	.27
Female	610	0	3	.84	.85	.70	.10	-.36	.20
Item 13									
Male	324	0	3	.74	.72	.78	.14	.54	.27
Female	614	0	3	.67	.76	.88	.10	.04	.20
Item 14									
Male	323	0	3	.71	.87	1.07	.14	.29	.27
Female	611	0	3	.78	.88	.96	.10	.16	.20
Item 15									
Male	322	0	3	.39	.65	1.75	.14	3.02	.27
Female	615	0	3	.37	.64	1.75	.10	2.61	.20
Item 16									
Male	323	0	3	.81	.85	.72	.14	-.40	.27
Female	612	0	3	.81	.86	.71	.10	-.43	.20
Item 17									
Male	323	0	3	.11	.40	4.05	.14	17.67	.27
Female	616	0	3	.42	.73	1.81	.10	2.78	.20
Item 18									
Male	324	0	3	.53	.77	1.45	.14	1.54	.27
Female	610	0	3	.79	.84	.89	.10	.15	.20
Item 19									
Male	323	0	3	.37	.65	1.79	.14	2.93	.27
Female	613	0	3	.39	.65	1.66	.10	2.43	.20
Item 20									
Male	323	0	3	.82	.89	.85	.14	-.11	.27
Female	613	0	3	.75	.84	.90	.10	.08	.20

Table 21. Descriptive Statistics of Item Responses for Males and Females on the One Item per Screen Format

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 1									
Male	323	0	3	.66	.78	1.11	.14	.84	.27
Female	615	0	3	.81	.78	.61	.10	-.29	.20
Item 2									
Male	321	0	3	.45	.70	1.57	.14	2.16	.27
Female	615	0	3	.56	.76	1.27	.10	1.07	.20
Item 3									
Male	318	0	3	.53	.75	1.29	.14	.97	.27
Female	609	0	3	.63	.80	1.15	.10	.67	.20
Item 4									
Male	323	0	3	.88	.93	.74	.14	-.45	.27
Female	611	0	3	.96	.93	.55	.10	-.74	.20
Item 5									
Male	322	0	3	1.42	.86	.14	.14	-.60	.27
Female	613	0	3	1.45	.87	.11	.10	-.65	.20
Item 6									
Male	323	0	3	.61	.85	1.26	.14	.67	.27
Female	613	0	3	.66	.87	1.20	.10	.60	.20
Item 7									
Male	324	0	3	1.27	.91	.25	.14	-.74	.27
Female	616	0	3	1.20	.93	.42	.10	-.67	.20
Item 8									
Male	321	0	3	1.06	.91	.42	.14	-.73	.27
Female	615	0	3	1.01	.91	.46	.10	-.76	.20
Item 9									
Male	324	0	3	.34	.67	2.14	.14	4.44	.27
Female	610	0	3	.28	.58	2.16	.10	4.41	0.20
Item 10									
Male	323	0	3	.17	.43	3.07	.14	12.29	.27
Female	614	0	3	.66	.83	1.06	.10	.26	.20

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 11									
Male	322	0	3	.94	.95	.73	.14	-.43	.27
Female	610	0	3	.93	.91	.68	.10	-.43	.20
Item 12									
Male	323	0	3	.82	.81	.68	.14	-.24	.27
Female	612	0	3	.83	.82	.70	.10	-.24	.20
Item 13									
Male	323	0	3	.80	.77	.78	.14	.30	.27
Female	614	0	3	.70	.77	.84	.10	-.01	.20
Item 14									
Male	323	0	3	.77	.91	.96	.14	-.06	.27
Female	614	0	3	.82	.90	.88	.10	-.09	.20
Item 15									
Male	321	0	3	.40	.63	1.63	.14	2.75	.27
Female	615	0	3	.38	.64	1.56	.10	1.65	.20
Item 16									
Male	323	0	3	.81	.84	.62	.14	-.62	.27
Female	613	0	3	.79	.84	.75	.10	-.33	.20
Item 17									
Male	322	0	3	.11	.41	3.99	.14	16.77	.27
Female	615	0	3	.42	.73	1.83	.10	2.90	.20
Item 18									
Male	323	0	3	.55	.78	1.40	.14	1.46	.27
Female	612	0	3	.83	.84	.79	.10	-.04	.20
Item 19									
Male	324	0	3	.41	.67	1.75	.14	2.98	.27
Female	613	0	3	.40	.64	1.51	.10	1.68	.20
Item 20									
Male	323	0	3	.83	.88	.83	.14	-.11	.27
Female	614	0	3	.74	.83	.93	.10	.19	.20

Table 22. Descriptive Statistics of Item Responses for Males and Females on the Multiple Items per Screen Format

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 1									
Male	324	0	3	.62	.75	1.14	.14	1.08	.27
Female	615	0	3	.78	.78	.72	.10	-.06	.20
Item 2									
Male	322	0	3	.47	.72	1.51	.14	1.74	.27
Female	608	0	3	.59	.79	1.20	.10	.67	.20
Item 3									
Male	318	0	3	.56	.78	1.31	.14	1.03	.27
Female	610	0	3	.65	.82	1.08	.10	.35	.20
Item 4									
Male	322	0	3	.93	.97	.68	.14	-.64	.27
Female	611	0	3	1.01	.95	.50	.10	-.82	.20
Item 5									
Male	323	0	3	1.40	.89	.12	.14	-.71	.27
Female	613	0	3	1.41	.90	.14	.10	-.75	.20
Item 6									
Male	324	0	3	.58	.85	1.45	.14	1.35	.27
Female	614	0	3	.67	.85	1.11	.10	.40	.20
Item 7									
Male	321	0	3	1.24	.92	.26	.14	-.77	.27
Female	616	0	3	1.15	.92	.46	.10	-.59	.20
Item 8									
Male	322	0	3	1.07	.92	.48	.14	-.65	.27
Female	611	0	3	1.08	.90	.39	.10	-.73	.20
Item 9									
Male	324	0	3	.33	.66	2.07	.14	3.91	.27
Female	612	0	3	.29	.59	2.15	.10	4.39	.20
Item 10									
Male	324	0	3	.20	.49	2.73	.14	8.39	.27
Female	614	0	3	.67	.81	1.06	.10	.45	.20

	n	Min	Max	Mean	*SD*	Skew	SE_{Skew}	Kurtosis	SE_{Kurt}
Item 11									
Male	324	0	3	.93	.97	.73	.14	-.53	.27
Female	616	0	3	.91	.92	.66	.10	-.56	.20
Item 12									
Male	320	0	3	.90	.85	.65	.14	-.28	.27
Female	608	0	3	.87	.87	.63	.10	-.53	.20
Item 13									
Male	323	0	3	.76	.76	.82	.14	.39	.27
Female	609	0	3	.72	.80	.85	.10	-.05	.20
Item 14									
Male	322	0	3	.75	.86	1.01	.14	.30	.27
Female	610	0	3	.78	.88	.93	.10	.05	.20
Item 15									
Male	324	0	3	.40	.67	1.81	.14	3.17	.27
Female	615	0	3	.40	.67	1.72	.10	2.61	.20
Item 16									
Male	323	0	3	.89	.90	.66	.14	-.55	.27
Female	611	0	3	.86	.89	.67	.10	-.54	.20
Item 17									
Male	324	0	2	.09	.33	4.14	.14	17.66	.27
Female	616	0	3	.43	.74	1.80	.10	2.69	.20
Item 18									
Male	324	0	3	0.52	0.74	1.36	0.14	1.36	0.27
Female	609	0	3	0.80	0.84	0.85	0.10	0.07	0.20
Item 19									
Male	320	0	3	0.42	0.67	1.53	0.14	1.70	0.27
Female	615	0	3	.43	.69	1.69	.10	2.65	.20
Item 20									
Male	323	0	3	.80	.87	.86	.14	-.11	.27
Female	614	0	3	.74	.84	.97	.10	.29	.20

Table 23. Results from Main Effects LMMs for Composite Scores

				NCP	
Variable	*df*	*F*	*p*	LL	UL
Sex	1,935	4.65	.031*	.22	14.46
English fluency	3,935	9.7	<.001**	12.28	47.42
Repeated Assessment	1,938	157.66	<.001**	117.52	203.51
Format	1,938	.16	.689	.00	3.85

Note. LL = lower limit of 90% CI of non-centrality parameter; UL = upper limit of 90% CI of non-centrality parameter.

$^{*}p < .05$. $^{**}p < .01$.

Table 24. Results from Main Effects LMMs for Subscale Scores

	Somatic Symptoms					Depressed Affect					Positive Affect					Interpersonal Problems				
				NCP					NCP					NCP					NCP	
Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
Sex	1,935	.07	.787	.00	2.92	1,935	22.41	<.001**	9.47	40.80	1,935	.009	.93	.00	0.83	1,935	.008	.929	.00	.71
English fluency	3,935	1.97	.116	.00	14.00	3,935	8.51	<.001**	9.88	42.68	3,935	21.05	<.001**	37.56	90.29	3,935	.318	.812	.00	3.20
Repeated Assessment	1,938	90.08	<.001**	60.91	124.82	1,938	80.51	<.001**	53.15	113.43	1,938	7.42	.007**	1.15	19.11	1,938	15.63	<.001**	5.29	31.41
Format	1,938	9.74	.002**	2.16	22.75	1,938	1.62	.204	.00	8.51	1,938	28.74	<.001**	13.71	49.24	1,938	1.35	.246	.00	7.88

Note. LL = lower limit of 90% CI of non-centrality parameter; UL = upper limit of 90% CI of non-centrality parameter.

$^{*}p < .05$. $^{**}p < .01$.

Table 25. Results from Main Effects LMMs for CES-D Items

	Item 1					Item 2					Item 3					Item 4				
				NCP					NCP					NCP					NCP	
Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
S	1,935	10.49	.001**	2.52	23.89	1,935	5.18	.023*	0.37	15.39	1,934	4.059	.044*	0.05	13.40	1,933	1.75	.187	.00	8.81
EF	3,935	4.08	.007**	1.95	24.06	3,934	.35	.791	.00	3.55	3,933	3.48	.016*	1.09	21.34	3,934	17.08	<.001**	28.37	75.64
RA	1,935	44.16	<.001**	24.79	69.02	1,925	8.31	.004**	1.52	20.53	1,918	29.21	<.001**	14.03	49.86	1,927	.28	.596	.00	4.57
F	1,935	5.72	.017*	.54	16.31	1,925	2.81	.094	.00	11.04	1,918	1.44	.23	.00	8.09	1,927	5.34	.021*	.42	15.66

	Item 5					Item 6					Item 7					Item 8				
				NCP					NCP					NCP					NCP	
	df	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
S	1,935	.15	.699	.00	3.77	1,935	1.68	.195	.00	8.65	1,936	1.98	.159	.00	9.32	1,935	.22	.639	.00	4.25
EF	3,935	.72	.539	.00	6.66	3,935	15.99	<.001**	1.95	24.06	3,936	8.23	<.001**	9.33	41.56	3,935	11.41	<.001**	15.85	54.12
RA	1,931	46.79	<.001**	26.76	72.32	1,933	30.4	<.001**	14.85	51.41	1,936	18.46	<.001**	6.98	35.38	1,929	4.36	.037*	.14	13.94
F	1,931	5.2	.023*	.38	15.42	1,933	.83	.361	.00	6.52	1,936	5.96	.015*	.62	16.71	1,929	6.19	.013*	.70	17.10

	Item 9					Item 10					Item 11					Item 12				
				NCP					NCP					NCP					NCP	
Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
S	1,935	1.94	.164	.00	9.23	1,935	105.29	<.001**	73.41	142.78	1,935	.04	0.852	.00	2.34	1,935	.05	.832	.00	2.57
EF	3,936	14.4	<.001**	22.35	65.57	3,935	.817	.485	.00	7.33	3,935	.79	.501	.00	7.15	3,935	19.93	<.001**	34.94	86.18
RA	1,932	6.31	.012*	.74	17.30	1,934	25.89	<.001**	11.77	45.47	1,931	1.41	.235	.00	8.02	1,925	1.03	.311	.00	7.07
F	1,932	.005	.946	.00	.24	1,934	1.1	.295	.00	7.25	1,931	.45	.504	.00	5.30	1,925	8.02	.005**	1.40	20.07
	Item 13					Item 14					Item 15					Item 16				
				NCP					NCP					NCP					NCP	
	df	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
S	1,934	1.7	.193	.00	8.70	1,933	.67	0.412	.00	6.04	1,936	.03	.875	.00	2.04	1,934	.278	.598	.00	4.55
EF	3,934	5.06	.002**	3.51	28.37	3,933	4.55	.004**	2.68	26.15	3,936	.38	.766	.00	3.86	3,935	14.85	<.001**	23.35	67.28
RA	1,929	20.1	<.001**	8.00	37.65	1,927	18.76	<.001**	7.17	35.80	1,935	3.35	.068	.00	12.08	1,930	4.6	.032*	.20	14.37
F	1,929	.05	.824	.00	2.57	1,927	4.35	.037*	.13	13.93	1,935	.26	.609	.00	4.47	1,930	14.64	<.001**	4.73	29.99

	Item 17					Item 18					Item 19					Item 20				
				NCP					NCP					NCP					NCP	
Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
S	1,935	60.38	<.001**	37.18	89.11	1,934	28.02	<.001**	13.21	48.29	1,936	0	1	.00	.00	1,935	1.92	.166	.00	9.19
EF	3,935	3.34	.019*	.90	20.69	3,934	3.86	.009**	1.63	23.07	3,935	.528	.663	.00	5.19	3,935	.4	.751	.00	4.05
RA	1,936	.67	.414	.00	6.04	1,928	5.79	.016*	.56	16.43	1,932	15.15	<.001**	5.02	30.72	1,933	.99	.319	.00	6.96
F	1,936	.18	.672	.00	3.99	1,928	4.75	.030*	.24	14.64	1,932	1.64	.201	.00	8.56	1,933	.49	.484	.00	5.44

Note. S = sex; EF = English fluency; RA = repeated assessment; F = format; LL = lower limit of 90% CI of non-centrality parameter; UL = upper limit of 90% CI of non-centrality parameter.

$^{*}p < .05.$ $^{**}p < .01.$

Table 26. Results from LMMs Including Terms of Interest and Appropriate Higher Order Interactions for Composite Scale Scores

					NCP	
	Variable	*df*	*F*	*p*	LL	UL
Terms addressing specific research questions	Sex	1,935	4.65	.031*	.22	14.46
	English fluency	3,935	9.7	<.001**	12.28	47.42
	Repeated Assessment	1,930	134.03	<.001**	97.44	176.28
	Format	1,930	.069	.793	.00	2.91
	S*RA	1,930	.953	.329	.00	6.86
	S*F	1,930	1.83	0.177	.00	8.99
	EF*RA	3,930	2.17	.091	.00	15.03
	EF*F	3,930	.17	.916	.00	1.13
Higher order interactions included in model	RA*F					
	S*RA*F					
	EF*RA*F					

Note. S = sex; EF = English fluency; RA = repeated assessment; F = format; LL = lower limit of 90% CI of non-centrality parameter; UL = upper limit of 90% CI of non-centrality parameter. Blank cells indicate the term was not included in the model for the scale scores.

$^{*}p < .05$. $^{**}p < .01$.

Table 27. Results from LMMs Including Terms of Interest and Appropriate Higher Order Interactions for Subscale Scores

		Somatic Symptoms					Depressed Affect					Positive Affect					Interpersonal Problems				
					NCP					NCP					NCP					NCP	
	Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
Terms addressing specific research questions	S	1,935	.07	.787	.00	2.92	1,935	22.41	<.001**	9.47	40.80	1,930	.11	.736	.00	3.42	1,931	0	.99	.00	.00
	EF	3,935	1.97	.116	.00	14.00	3,935	8.51	<.001**	9.88	42.68	3,930	21.71	<.001**	39.11	.00	3,931	.25	.863	.00	2.37
	RA	1,930	61.31	<.001**	37.90	90.25	1,930	68.8	<.001**	43.79	99.35	1,930	12.55	<.001**	3.58	26.96	1,930	13.32	<.001**	3.99	28.08
	F	1,930	9.85	.002**	2.22	22.91	1,930	1.22	.269	.00	7.56	1,930	15.44	<.001**	5.18	31.14	1,930	.79	.374	.00	6.40
	S*RA	1,930	.11	.737	.00	3.42	1,930	1.19	.276	.00	7.48	1,930	1.09	.297	.00	7.23	1,930	.01	.945	.00	.93
	S*F	1,930	.72	.397	.00	6.20	1,930	.82	.366	.00	6.49	1,930	.09	.772	.00	3.20	1,930	.71	.400	.00	6.17
	EF*RA	3,930	2.43	.064	.00	16.34	3,930	.94	.422	.00	3.00	3,930	2.21	.085	.00	15.23	3,931	1.6	.195	.00	12.03
	EF*F	3,930	.3	.829	.00	3.00	3,930	.27	.851	.00	2.63	3,930	.34	.8	.00	3.44	3,931	.74	.529	.00	6.80
Higher order interactions included in model	RA*F											1,930	3.15	.076	.00	11.70	1,931	3.76	.053	.00	12.85
	S*RA*F											1,930	2.47	.117	.00	10.35					
	EF*RA*F											3,930	2.13	.095	.00	14.83	3,931	3.62	.013*	1.28	21.98

Note. S = sex; EF = English fluency; RA = repeated assessment; F = format; LL = lower limit of 90% CI of non-centrality parameter; UL = upper limit of 90% CI of non-centrality parameter. Blank cells indicate the term was not included in the model for the subscale.

$^{*}p < .05$. $^{**}p < .01$.

Table 28. Results from LMMs Including Terms of Interest and Appropriate Higher Order Interactions for each CES-D Item

		Item 1					Item 2					Item 3				
					NCP					NCP					NCP	
	Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
Terms addressing specific research questions	S	1,935	10.5	.001**	2.53	23.90	1,935	5.18	.023*	.37	15.39	1,934	4.05	.044*	.05	13.39
	EF	3,935	4.07	.007**	1.94	24.02	3,934	.35	.79	.00	3.55	3,933	3.48	.016*	1.09	21.34
	RA	1,927	28.98	<.001**	13.87	49.55	1,917	4.98	.026*	.31	15.04	1,910	21.73	<.001**	9.03	39.88
	F	1,927	8.39	.004**	1.56	1.56	1,917	1.73	.189	.00	8.76	1,910	2.15	.143	.00	9.68
	S*RA	1,927	.8	.371	.00	6.43	1,917	.1	.747	.00	3.31	1,911	2.52	.113	.00	10.45
	S*F	1,927	.09	.762	.00	3.20	1,917	.03	.851	.00	2.04	1,911	.06	.805	.00	2.76
	EF*RA	3,927	1.47	.221	.00	11.31	3,917	2.62	.049*	.01	17.27	3,910	.14	.938	.00	.54
	EF*F	3,927	1.41	.238	.00	10.97	3,917	1.26	.287	.00	10.11	3,910	.44	.721	.00	4.43
Higher order interactions included in model	RA*F															
	S*EF															
	S*RA*F															
	EF*RA*F															
	S*EF*RA															
	S*EF*F															

		Item 4					Item 5					Item 6				
					NCP					NCP					NCP	
	Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
Terms addressing specific research questions	S	1,928	2.62	.106	.00	10.66	1,932	.01	.919	.00	.93	1,935	1.69	.194	.00	8.67
	EF	3,929	18.01	$<.001^{**}$	30.50	79.10	3,932	.63	.598	.00	6.00	3,935	15.98	$<.001^{**}$	25.88	71.53
	RA	1,918	2.84	.092	.00	11.09	1,920	21.38	$<.001^{**}$	8.81	39.40	1,925	50.02	$<.001^{**}$	29.20	76.34
	F	1,918	1.44	.231	.00	8.09	1,920	4.53	$.034^{*}$	.18	14.25	1,925	2.49	0.115	0.00	10.39
	S*RA	1,918	0.82	.367	.00	6.49	1,920	1.16	.282	.00	7.41	1,925	4.38	$.037^{*}$	0.14	13.98
	S*F	1,918	0	.958	.00	.00	1,920	1.53	.217	.00	8.31	1,925	1.82	0.178	0.00	8.97
	EF*RA	3,919	1.4	.242	.00	10.92	3,920	1.53	.205	.00	11.65	3,925	5.01	$.002^{**}$	3.43	28.15
	EF*F	3,919	1.97	.117	.00	14.00	3,920	.43	.733	.00	4.34	3,925	0.85	0.467	0.00	7.56
Higher order interactions included in model	RA*F	1,928	1.83	0.177	0.00	8.99										
	S*EF						3,932	0.24	0.868	0.00	2.24					
	S*RA*F	1,928	4.05	$.045^{*}$	0.05	13.39										
	EF*RA*F	3,929	2.21	0.085	0.00	15.23										
	S*EF*RA															
	S*EF*F						3,920	2.99	$.030^{*}$	0.45	19.05					

		Item 7					Item 8					Item 9				
					NCP					NCP					NCP	
	Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
Terms addressing specific research questions	S	1,936	1.99	.159	.00	9.34	1,935	.218	.64	.00	4.24	1,935	1.94	0.164	0.00	9.23
	EF	3,936	8.24	<.001**	9.35	41.60	3,935	11.42	<.001**	15.87	54.16	3,936	14.39	<.001**	22.32	65.54
	RA	1,929	14.55	<.001**	4.67	29.86	1,921	4.62	.032*	0.21	14.41	1,923	5.07	.025*	0.34	15.20
	F	1,929	2.3	0.13	0.00	10.00	1,921	1.56	0.212	0.00	8.38	1,923	0.04	0.845	0.00	2.34
	S*RA	1,929	0.02	0.894	0.00	1.63	1,921	0.54	0.462	0.00	5.62	1,923	0.42	0.518	0.00	5.18
	S*F	1,929	0.4	0.528	0.00	5.10	1,921	2.11	0.147	0.00	9.60	1,923	0.66	0.417	0.00	6.01
	EF*RA	3,929	0.19	0.903	0.00	1.48	3,921	1.07	0.361	0.00	8.97	3,923	0.98	0.4	0.00	8.40
	EF*F	3,929	0.25	0.864	0.00	2.37	3,921	0.28	0.844	0.00	2.76	3,923	1.72	0.161	0.00	12.68
Higher order interactions included in model	RA*F															
	S*EF															
	S*RA*F															
	EF*RA*F															
	S*EF*RA															
	S*EF*F															

		Item 10					Item 11					Item 12				
					NCP					NCP					NCP	
	Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
Terms addressing specific research questions	S	1,935	105.21	<.001**	73.34	142.68	1,935	0.04	0.853	0.00	2.34	1,933	0	0.953	0.00	0.00
	EF	3,935	0.82	0.483	0.00	7.36	3,935	0.79	0.5	0.00	7.15	3,933	20.26	<.001**	35.71	87.40
	RA	1,926	14.21	<.001**	4.48	29.37	1,923	1.02	0.314	0.00	7.04	1,917	1.12	0.29	0.00	7.30
	F	1,926	1.99	0.159	0.00	9.34	1,923	0.8	0.372	0.00	6.43	1,917	7.06	.008**	1.01	18.53
	S*RA	1,926	0.44	0.507	0.00	5.26	1,923	0	0.98	0.00	0.00	1,917	0.2	0.653	0.00	4.13
	S*F	1,926	0.71	0.398	0.00	6.17	1,923	0.01	0.937	0.00	0.93	1,917	1.18	0.277	0.00	7.46
	EF*RA	3,926	0.77	0.513	0.00	7.01	3,923	0.39	0.763	0.00	3.96	3,918	0.3	0.825	0.00	3.00
	EF*F	3,926	0.43	0.733	0.00	4.34	3,923	0.25	0.858	0.00	2.37	3,918	0.35	0.788	0.00	3.55
Higher order interactions included in model	RA*F											1,933	2.62	0.106	0.00	10.66
	S*EF															
	S*RA*F											1,933	2.52	0.113	0.00	10.45
	EF*RA*F															
	S*EF*RA															
	S*EF*F															

		Item 13					Item 14					Item 15				
					NCP					NCP					NCP	
	Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
Terms addressing specific research questions	Sex	1,934	1.69	0.193	0.00	8.67	1,930	1.19	0.276	0.00	7.48	1,930	0	0.951	0.00	0.00
	EF	3,933	5.06	.002**	3.51	28.37	3,930	4.12	.007**	2.01	24.24	3,929	0.26	0.858	0.00	2.51
	RA	1,910	23.4	<.001**	10.12	42.14	1,916	17.08	<.001**	6.15	33.46	1,925	1.72	0.19	0.00	2.51
	F	1,910	0.36	0.548	0.00	4.94	1,916	3.54	0.06	0.00	12.44	1,925	0.22	0.639	0.00	4.25
	S*RA	1,911	0.01	0.91	0.00	0.93	1,915	3.42	0.065	0.00	12.22	1,925	3.66	0.056	0.00	12.67
	S*F	1,911	3.7	0.055	0.00	12.74	1,916	3.15	0.076	0.00	11.70	1,924	0.3	0.582	0.00	4.67
	EF*RA	3,910	1.92	0.125	0.00	13.74	3,916	0.37	0.772	0.00	3.76	3,925	2.17	0.09	0.00	15.03
	EF*F	3,910	0.1	0.958	0.00	0.00	3,916	0.84	0.47	0.00	7.49	3,925	1.81	0.144	0.00	13.16
Higher order interactions included in model	RA*F											1,930	7.04	.008**	1.01	18.49
	S*EF						3,930	0.54	0.657	0.00	5.29	3,929	0.01	0.998	0.00	0.00
	S*RA*F															
	EF*RA*F											3,929	3.33	.019*	0.89	20.64
	S*EF*RA											3,925	2.67	.047*	0.06	17.51
	S*EF*F						3,916	5.18	.001**	3.71	28.89					

		Item 16					Item 17					Item 18				
					NCP					NCP					NCP	
	Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
Terms addressing specific research questions	S	1,934	0.28	0.599	0.00	4.57	1,936	60.33	<.001**	37.14	89.04	1,934	28.01	<.001**	13.21	48.28
	EF	3,935	14.84	<.001**	23.32	67.24	3,935	3.35	.019*	0.91	20.74	3,934	3.86	.009**	1.63	23.07
	RA	1,921	6.15	.013*	0.68	17.03	1,928	0.31	0.577	0.00	4.72	1,919	3.97	.047*	0.03	13.24
	F	1,921	11.35	.001**	2.95	25.18	1,928	1.57	0.211	0.00	8.40	1,919	2.53	0.112	0.00	10.47
	S*RA	1,921	2.37	0.124	0.00	10.14	1,928	0.4	0.528	0.00	5.10	1,919	0.3	0.587	0.00	4.67
	S*F	1,921	0.06	0.803	0.00	2.76	1,928	1.87	0.172	0.00	9.08	1,919	0.01	0.922	0.00	0.93
	EF*RA	3,922	0.72	0.54	0.00	6.66	3,928	0.79	0.501	0.00	7.15	3,919	1.23	0.3	0.00	9.93
	EF*F	3,922	1.08	0.357	0.00	9.03	3,928	0.54	0.658	0.00	5.29	3,919	0.25	0.86	0.00	2.37
Higher order interactions included in model	RA*F															
	S*EF															
	S*RA*F															
	EF*RA*F															
	S*EF*RA															
	S*EF*F															

		Item 19					Item 20				
					NCP					NCP	
	Variable	*df*	*F*	*p*	LL	UL	*df*	*F*	*p*	LL	UL
Terms addressing specific research questions	S	1,932	0.01	0.944	0.00	0.93	1,935	1.92	0.166	0.00	9.19
	EF	3,931	0.54	0.653	0.00	5.29	3,935	0.4	0.751	0.00	4.05
	RA	1,924	13.38	<.001**	4.02	28.17	1,924	0.75	0.387	0.00	6.29
	F	1,924	0.56	0.454	0.00	5.69	1,924	1.16	0.282	0.00	7.41
	S*RA	1,924	1.5	0.221	0.00	8.24	1,925	0.2	0.653	0.00	4.13
	S*F	1,924	0.87	0.351	0.00	6.63	1,925	0.6	0.439	0.00	5.82
	EF*RA	3,924	1.68	0.169	0.00	12.47	3,925	1.24	0.293	0.00	9.99
	EF*F	3,924	0.96	0.409	0.00	8.28	3,925	1.85	0.136	0.00	13.38
Higher order interactions included in model	RA*F										
	S*EF	1,931	0.64	0.425	0.00	5.95					
	S*RA*F										
	EF*RA*F	3,931	2.74	.042*	0.15	17.85					
	S*EF*RA										
	S*EF*F										

Note. S = sex; EF = English fluency; RA = repeated assessment; F = format; LL = lower limit of 90% CI of non-centrality parameter; UL = upper limit of 90% CI of non-centrality parameter. Blank cells indicate the term was not included in the model for the item.

$^{*}p < .05$. $^{**}p < .01$.

Table 29. Results from GEE Main Effects Models of Categorizations of CES-D Scores

	2-group categorization (<16)					3-group categorization (>=16 and <=23)				
				NCP					NCP	
Variable	*df*	*Wald* χ^2	*p*	LL	UL	*df*	*Wald* χ^2	*p*	LL	UL
S	1	5.62	.018*	0.51	16.12	1	5.33	.021*	0.42	15.63
EF	3	30.44	<.001**	13.32	49.01	3	33.3	<.001**	15.33	52.72
RF	1	25.98	<.001**	11.92	45.45	1	18.78	<.001**	7.23	35.74
F	1	1.48	0.223	0.00	8.19	1	1.38	0.239	0.00	7.95

Note. S = sex; EF = English fluency; RA = repeated assessment; F = format; LL = lower limit of 90% CI of non-centrality parameter; UL = upper limit of 90% CI of non-centrality parameter.

$^{*}p < .05$. $^{**}p < .01$.

Table 30. Results from CES-D Categorization GEE Models Including All Terms of Interest

	2-group categorization (<16)					3-group categorization (>=16 and <=23)				
				NCP					NCP	
Variable	*df*	*Wald* χ^2	*p*	LL	UL	*df*	*Wald* χ^2	*p*	LL	UL
S	1	5.47	.019[*]	.46	15.87	1	5.31	.021[*]	5.76	15.60
EF	3	30.08	<.001[**]	13.07	48.54	3	33.27	<.001[**]	15.31	52.68
RA	1	20.09	<.001[**]	8.05	37.54	1	14.26	<.001[**]	4.54	29.39
F	1	2.77	.096	.00	10.95	1	2.09	.149	.00	9.55
S*RA	1	.48	.49	.00	5.41	1	.11	.738	.00	3.42
S*F	1	.04	.838	.00	2.34	1	.19	.66	.00	4.06
EF*RA	3	1.8	.615	.00	5.76	3	1.69	.638	.00	5.47
EF*F	3	4.76	.191	.00	11.95	3	4.93	.177	.00	12.26

Note. S = sex; EF = English fluency; RA = repeated assessment; F = format; LL = lower limit of 90% CI of non-centrality parameter; UL = upper limit of 90% CI of non-centrality parameter.

[*]p < .05. [**]p < .01.

Table 31. Number and Percentage of Participants in each CES-D Category on the One Item and Multiple Items per Screen Formats

	One item *n* (%)	Multiple items *n* (%)
2-group categorization		
Less than 16 (<16)	590 (62.8)	576 (61.3)
Greater than or equal to 16 (≥16)	350 (37.2)	364 (38.7)
3-group categorization		
Less than 16 (<16)	590 (62.8)	576 (61.3)
Between 16 and 23 (≥16 and ≤23)	185 (19.7)	197 (21.0)
Greater than 23 (>23)	165 (17.6)	167 (17.8)

Note. N=940

Table 32. Model Fit for Confirmatory and Multigroup Factor Analysis for Males and Females on Each Assessment

		Model χ^2			RMSEA			SRMR	χ^2 Difference Test		
		df	χ^2	*p*	RMSEA	LL	UL		*df*	χ^2	*p*
First Assessment	CFA male	164	311.33	<.001	.053	.044	.062	.051			
	CFA female	164	439.18	<.001	.052	.046	.058	.044			
	Configural invariance	328	751.26	<.001	.052	.047	.057	.047			
	Weak factorial invariance	344	901.61	<.001	.059	.054	.063	.068	16	128.21	<.001[a]
	Strong factorial invariance	364	1161.77	<.001	.068	.064	.073	.835	20	285.49	<.001[b]
Second Assessment	CFA male	164	368.35	<.001	.062	.054	.070	.059			
	CFA female	164	551.82	<.001	.062	.056	.068	.051			
	Configural invariance	328	913.06	<.001	.062	.057	.067	.054			
	Weak factorial invariance	344	1064.52	<.001	.067	.062	.071	.075	16	118.65	<.001[a]
	Strong factorial invariance	364	1288.83	<.001	.074	.069	.078	.086	20	248.13	<.001[b]

Note. LL = lower limit of 90% CI of RMSEA values; UL = upper limit of 90% CI of RMSEA values.

[a] Difference test comparing weak factorial invariance and configural invariance. [b] Difference test comparing strong factorial invariance and weak factorial invariance.

Table 33. Model Fit for Confirmatory and Multigroup Factor Analysis for Males and Females on Each Assessment when Items 10 and 17 are Dropped from the Model

		Model χ^2			RMSEA			SRMR	χ^2 Difference Test		
		df	χ^2	*p*	RMSEA	LL	UL		*df*	χ^2	*p*
First Assessment	CFA male	129	203.23	<.001	.042	.031	.053	.043			
	CFA female	129	269.29	<.001	.042	.035	.049	.040			
	Configural invariance	258	472.94	<.001	.042	.036	.048	.041			
	Weak factorial invariance	272	510.44	<.001	.043	.037	.049	.049	14	36.72	<.001 [a]
	Strong factorial invariance	290	592.31	<.001	.047	.042	.053	.051	18	88.12	<.001 [b]
Second Assessment	CFA male	129	246.95	<.001	.053	.043	.063	.051			
	CFA female	129	362.36	<.001	.053	.048	.061	.048			
	Configural invariance	258	609.43	<.001	.054	.048	.059	.049			
	Weak factorial invariance	272	640.30	<.001	.054	.048	.059	.055	14	31.26	.005 [a]
	Strong factorial invariance	290	719.36	<.001	.056	.051	.061	.057	18	88.46	<.001 [b]

Note. LL = lower limit of 90% CI of RMSEA values; UL = upper limit of 90% CI of RMSEA values.

[a] Difference test comparing weak factorial invariance and configural invariance. [b] Difference test comparing strong factorial invariance and weak factorial invariance.

Table 34. Follow-up Pairwise Comparisons for Composite and Subscale Scores where there was a Main Effect of English Fluency in the LMM Model

	Comparison		Mean Difference	*df*	p_{mc}
Composite score	Very fluent, English is my first language	More fluent than my first language	-1.46	935	.265
Theoretical range: 0-60		Same fluency as my first language	-2.82	935	.008**
		Less fluent than my first language	-4.64	935	<.001**
	More fluent than my first language	Same fluency as my first language	-1.36	935	.990
		Less fluent than my first language	-3.18	935	.015*
	Same fluency as my first language	Less fluent than my first language	-1.82	935	.691
Depressed affect	Very fluent, English is my first language	More fluent than my first language	-.21	935	>.99
Theoretical range: 0-21		Same fluency as my first language	-.94	935	.078
		Less fluent than my first language	-1.94	935	<.001**
	More fluent than my first language	Same fluency as my first language	-.73	935	.495
		Less fluent than my first language	-1.73	935	.001**
	Same fluency as my first language	Less fluent than my first language	-1.00	935	.269
Positive affect	Very fluent, English is my first language	More fluent than my first language	-1.13	935	<.001**
Theoretical range: 0-12		Same fluency as my first language	-1.32	935	<.001**
		Less fluent than my first language	-1.86	935	<.001**
	More fluent than my first language	Same fluency as my first language	-.19	935	>.99
		Less fluent than my first language	-.73	935	.139
	Same fluency as my first language	Less fluent than my first language	-.54	935	.773

Note. Mean difference is based estimated marginal means from the LMM model with sex, English fluency, format, and repeated assessment as factors. Negative mean difference scores indicate the group with the lower level of English fluency has higher reported depressive symptomology. p_{mc} has been adjusted using a Bonferroni correction.

*p<.05. **p<.01.

Table 35. Number and Percentage of Participants in each CES-D Category for Males and Females

	Males	Females
	n (%)	*n* (%)
2-group categorization		
Less than 16 (<16)	210 (64.8)	350 (56.8)
Greater than or equal to 16 (≥16)	114 (35.2)	266 (43.2)
3-group categorization		
Less than 16 (<16)	210 (64.8)	350 (56.8)
Between 16 and 23 (≥16 and ≤23)	63 (19.4)	138 (22.4)
Greater than 23 (>23)	51 (15.7)	128 (20.8)

Note. $n_{\text{male}} = 324$; $n_{\text{female}} = 616$.

Table 36. Follow-up Pairwise Comparisons for Item Responses where there was a Main Effect of English Fluency in the LMM Model

Item	Comparison		Mean Difference	*df*	p_{mc}
Item 1	Very fluent, English is my first language	More fluent than my first language	-.019	934	>.99
		Same fluency as my first language	-.075	934	>.99
		Less fluent than my first language	-.263	936	.004**
	More fluent than my first language	Same fluency as my first language	-.056	934	>.99
		Less fluent than my first language	-.243	935	.024*
	Same fluency as my first language	Less fluent than my first language	-.188	935	.266
Item 4	Very fluent, English is my first language	More fluent than my first language	-.341	934	<.001**
		Same fluency as my first language	-.424	935	<.001**
		Less fluent than my first language	-.479	933	<.001**
	More fluent than my first language	Same fluency as my first language	-.083	935	<.001**
		Less fluent than my first language	-.138	933	<.001**
	Same fluency as my first language	Less fluent than my first language	-.055	934	<.001**
Item 7	Very fluent, English is my first language	More fluent than my first language	-.162	935	.119
		Same fluency as my first language	-.326	937	.001**
		Less fluent than my first language	-.341	935	.001**
	More fluent than my first language	Same fluency as my first language	-.164	937	.489
		Less fluent than my first language	-.178	935	.460
	Same fluency as my first language	Less fluent than my first language	-.015	936	>.99
Item 8	Very fluent, English is my first language	More fluent than my first language	-.356	935	<.001**
		Same fluency as my first language	-.205	933	.081
		Less fluent than my first language	-.344	936	.001**
	More fluent than my first language	Same fluency as my first language	.151	934	.624
		Less fluent than my first language	.011	936	>.99
	Same fluency as my first language	Less fluent than my first language	-.139	935	>.99
Item 9	Very fluent, English is my first language	More fluent than my first language	-.092	936	.278
		Same fluency as my first language	-.222	936	<.001**
		Less fluent than my first language	-.357	936	<.001**
	More fluent than my first language	Same fluency as my first language	-.130	936	.213
		Less fluent than my first language	-.265	936	<.001**
	Same fluency as my first language	Less fluent than my first language	-.135	936	.396

Item	Comparison		Mean Difference	*df*	p_{mc}
Item 12	Very fluent, English is my first language	More fluent than my first language	-.222	936	.002**
		Same fluency as my first language	-.348	935	<.001**
		Less fluent than my first language	-.568	935	<.001**
	More fluent than my first language	Same fluency as my first language	-.126	935	.822
		Less fluent than my first language	-.346	936	.001**
	Same fluency as my first language	Less fluent than my first language	-.220	935	.167
Item 13	Very fluent, English is my first language	More fluent than my first language	-.018	935	>.99
		Same fluency as my first language	-.161	935	0.149
		Less fluent than my first language	-.268	933	0.004**
	More fluent than my first language	Same fluency as my first language	-.143	935	0.444
		Less fluent than my first language	-.250	933	0.022*
	Same fluency as my first language	Less fluent than my first language	-.107	934	1.00
Item 16	Very fluent, English is my first language	More fluent than my first language	-.220	934	.005**
		Same fluency as my first language	-.354	936	<.001**
		Less fluent than my first language	-.477	933	<.001**
	More fluent than my first language	Same fluency as my first language	-.134	936	.766
		Less fluent than my first language	-.258	933	.039*
	Same fluency as my first language	Less fluent than my first language	-.123	934	>.99
Item 18	Very fluent, English is my first language	More fluent than my first language	-.019	935	>.99
		Same fluency as my first language	-.179	935	.109
		Less fluent than my first language	-.225	933	.038*
	More fluent than my first language	Same fluency as my first language	-.160	935	.352
		Less fluent than my first language	-.206	934	.139
	Same fluency as my first language	Less fluent than my first language	-.046	934	>.99

Note. Mean difference is based estimated marginal means from the LMM model with sex, English fluency, format, and repeated assessment as factors. Negative mean difference scores indicate the group with the lower level of English fluency has higher reported depressive symptomology. p_{mc} has been adjusted using a Bonferroni correction. Theoretical range for items: 0-3.

*p<.05. **p<.01

Table 37. Model Fit for Confirmatory and Multigroup Factor Analysis for Participants who Identified as Very Fluent in English and Participants who did not Identify as Very Fluent in English on the First and Second Assessment

		Model χ^2			RMSEA			SRMR	χ^2 Difference Test		
		df	χ^2	*p*	RMSEA	LL	UL		*df*	χ^2	*p*
First Assessment	CFA very fluent, English is my first language	164	438.17	<.001	.059	.053	.066	.052			
	CFA English not first language	164	459.64	<.001	.062	.056	.069	.051			
	Configural invariance	328	897.81	<.001	.061	.056	.066	.052			
	Weak factorial invariance	344	929.03	<.001	.060	.056	.065	.059	16	32.68	.008 [a]
	Strong factorial invariance	364	1072.55	<.001	.064	.060	.069	.069	20	156.96	<.001 [b]
Second Assessment	CFA very fluent, English is my first language	164	505.36	<.001	.066	.060	.073	.059			
	CFA English not first language	164	558.25	<.001	.072	.065	.078	.060			
	Configural invariance	328	1063.32	<.001	.069	.064	.074	.060			
	Weak factorial invariance	344	1102.52	<.001	.068	.064	.073	.068	16	41.13	<.001 [a]
	Strong factorial invariance	364	1214.35	<.001	.071	.066	.075	.075	20	122.71	<.001 [b]

Note. LL = lower limit of 90% CI of RMSEA values; UL = upper limit of 90% CI of RMSEA values.

[a] Difference test comparing weak factorial invariance and configural invariance. [b] Difference test comparing strong factorial invariance and weak factorial invariance.

Table 38. Model Fit for Confirmatory and Multigroup Factor Analysis for Participants who Identified as Very Fluent in English and Participants who did not Identify and Very Fluent in English on the First and Second Assessment when Items 10 and 17 are Dropped

		Model χ^2			RMSEA			SRMR	χ^2 Difference Test		
		df	χ^2	*p*	RMSEA	LL	UL		*df*	χ^2	*p*
First Assessment	CFA very fluent, English is my first language	129	235.03	<.001	.042	.033	.050	.040			
	CFA English not first language	129	245.84	<.001	.044	.036	.052	.044			
	Configural invariance	258	480.79	<.001	.043	.037	.049	.042			
	Weak factorial invariance	272	518.50	<.001	.044	.038	.050	.052	14	37.08	<.001 [a]
	Strong factorial invariance	290	645.18	<.001	.051	.046	.056	.065	18	138.53	<.001 [b]
Second Assessment	CFA very fluent, English is my first language	129	302.41	<.001	.053	.045	.061	.049			
	CFA English not first language	129	351.61	<.001	.061	.053	.069	.055			
	Configural invariance	258	653.64	<.001	.057	.052	.063	.052			
	Weak factorial invariance	272	698.49	<.001	.058	.052	.063	.063	14	43.71	<.001 [a]
	Strong factorial invariance	290	796.81	<.001	.061	.056	.066	.072	18	107.74	<.001 [b]

Note. LL = lower limit of 90% CI of RMSEA values; UL = upper limit of 90% CI of RMSEA values.

[a] Difference test comparing weak factorial invariance and configural invariance. [b] Difference test comparing strong factorial invariance and weak factorial invariance.

Table 39. Number and Percentage of Participants in each CES-D Category for Participants in Each Self-rated English Fluency Group

	VF *n* (%)	MF *n* (%)	SF *n* (%)	LF *n* (%)
2-group categorization				
Less than 16 (<16)	319 (67.2)	129 (57.8)	67 (50.0)	45 (41.7)
Greater than or equal to 16 (>=16)	156 (32.8)	94 (42.2)	67 (50.0)	63 (58.3)
3-group categorization				
Less than 16 (<16)	319 (67.2)	129 (57.8)	67 (50.0)	45 (41.7)
Between 16 and 23 (>=16 and <=23)	84 (17.7)	51 (22.9)	35 (26.1)	31 (28.7)
Greater than 23 (>23)	72 (15.2)	43 (19.3)	32 (23.9)	32 (29.6)

Note. $n_{VF} = 475$; $n_{MF} = 223$; $n_{SF} = 134$; $n_{LF} = 108$. VF = very fluent, English is my first language; MF = more fluent in English than my first language; SF = same fluency in English as my first language; LF = less fluent in English than first my language.

Table 40. Model Fit for Confirmatory and Multigroup Factor Analysis for the One Item per Screen Format and Multiple Items per Screen Format on the First and Second Assessment

		Model χ^2			RMSEA			SRMR	χ^2 Difference Test		
		df	χ^2	*p*	RMSEA	LL	UL		*df*	χ^2	*p*
First Assessment	CFA multiple items per screen	164	438.86	<.001	.058	.051	.064	.048			
	CFA one item per screen	164	473.59	<.001	.066	.059	.073	.055			
	Configural invariance	328	911.95	<.001	.062	.057	.066	.052			
	Weak factorial invariance	344	913.24	<.001	.059	.055	.064	.053	16	6.39	.983[a]
	Strong factorial invariance	364	941.49	<.001	.058	.054	.063	.054	20	23.97	.244[b]
Second Assessment	CFA multiple items per screen	164	505.98	<.001	.069	.062	.076	.059			
	CFA one item per screen	164	530.34	<.001	.067	.060	.073	.057			
	Configural invariance	328	1037.09	<.001	.068	.063	.072	.058			
	Weak factorial invariance	344	1055.07	<.001	.066	.062	.071	.061	16	22.53	.127[a]
	Strong factorial invariance	364	1087.36	<.001	.065	.061	.069	.062	20	40.54	.004[b]

Note. LL = lower limit of 90% CI of RMSEA values; UL = upper limit of 90% CI of RMSEA values.

[a] Difference test comparing weak factorial invariance and configural invariance. [b] Difference test comparing strong factorial invariance and weak factorial invariance.

Table 41. **Model Fit for Confirmatory and Multigroup Factor Analysis for the One Item per Screen Format and Multiple Items per Screen Format on the Second Assessment when the Intercepts for Items 5 and 7 are not Constrained in the Test of Strong Factorial Invariance**

		Model χ^2			RMSEA			SRMR	χ^2 Difference Test		
		df	χ^2	*p*	RMSEA	LL	UL		*df*	χ^2	*p*
Second Assessment	CFA multiple items per screen	164	505.98	<.001	.069	.062	.076	.059			
	CFA one item per screen	164	530.34	<.001	.067	.060	.073	.057			
	Configural invariance	328	1037.09	<.001	.068	.063	.072	.058			
	Weak factorial invariance	344	1055.07	<.001	.066	.062	.071	.061	16	22.53	.127[a]
	Strong factorial invariance	362	1079.36	<.001	.065	.060	.069	.061	18	16.75	.540[b]

Note. LL = lower limit of 90% CI of RMSEA values; UL = upper limit of 90% CI of RMSEA values.

[a] Difference test comparing weak factorial invariance and configural invariance. [b] Difference test comparing strong factorial invariance and weak factorial invariance.

Table 42. Number and Percentage of Participants in each CES-D Category on the First Assessment and Second Assessment

	First presentation	Second presentation
	n (%)	*n* (%)
2-group categorization		
Less than 16 (<16)	560 (59.6)	606 (64.5)
Greater than or equal to 16 (≥16)	380 (40.4)	334 (35.5)
3-group categorization		
Less than 16 (<16)	560 (59.6)	606 (64.5)
Between 16 and 23 (≥16 and ≤23)	201 (21.4)	181 (19.3)
Greater than 23 (>23)	179 (19.0)	153 (16.3)

Note. *N*=940.

Table 43. Mean Item Response for Item 6, “I felt depressed”, for Participants at each Self-rated English Fluency Level

		First presentation		Second Presentation	
	n	Mean	*SD*	Mean	*SD*
Very fluent, English is my first language	472	.52	.80	.49	.79
More fluent than my first language	222	.70	.83	.61	.84
Same fluency as my first language	132	.81	.85	.68	.85
Less fluent than my first language	108	1.18	.95	.98	1.01

Note. Range of possible responses to item 6: 0-3.

Table 44. Model Fit for Confirmatory Factor Models and to Modified Models used to Assess Measurement Invariance Across Repeated Assessment

	Model χ^2			RMSEA			SRMR	χ^2 Difference Test		
	df	χ^2	*p*	RMSEA	LL	UL		*df*	χ^2	*p*
CFA first assessment	164	689.25	<.001	.058	.054	.063	.046			
CFA second assessment	164	833.97	<.001	.066	.062	.070	.053			
Model 1: not constrained [a]	714	3516.38	<.001	.046	.044	.047	.052			
Model 2: factor loadings constrained [b]	734	3604.72	<.001	.046	.044	.047	.052	20	89.31	<.001 [e]
Model3a: intercepts partially constrained [c]	744	3664.82	<.001	.046	.044	.047	.052	10	62.67	<.001 [f]
Model 3b: intercepts constrained [d]	754	4077.18	<.001	.048	.047	.050	.053	11	499.15	<.001 [g]

Note. The structure of models 1, 2, and 3 is presented in Figure 5. LL = lower limit of 90% CI of RMSEA values; UL = upper limit of 90% CI of RMSEA values. [a] item factor loadings and intercepts were not constrained. [b] factor loading for each item on the first assessment was constrained as equal to factor loading of the corresponding item on the second assessment. [c] in addition to constrained factor loadings, the intercept of an item on the first assessment was constrained as equal to the intercept of the corresponding item on the second assessment for the 10 items where no mean differences were identified due to repeated assessment in LMM analysis. [d] in addition to constrained factor loadings, the intercept of an item on the first assessment was constrained as equal to the intercept of the corresponding item for each item. [e] Difference test comparing model 2 to model 1. [f] Difference test comparing model 3a to model 2. [g] Difference test comparing model 3b to model 3a.

Table 45. List of CES-D Items and Coding of Item Characteristics used in Research Question 4

Item	Flesch-Kincaid	Reverse Coded	Refers to others	Refers to perceptions of others	Behaviour (B) Feeling (F)	Somatic Subscale	Subscale
1. I was bothered by things that usually don't bother me.	4.8	N	N	N	F	Y	Som
2. I did not feel like eating; my appetite was poor.	1.7	N	N	N	B	Y	Som
3. I felt that I could not shake off the blues even with help from my family or friends.	5.1	N	Y	N	F	N	Dep
4. I felt I was just as good as other people.	2.4	Y	Y	N	F	N	Pos
5. I had trouble keeping my mind on what I was doing.	2.6	N	N	N	B	Y	Som
6. I felt depressed.	1.3	N	N	N	F	N	Dep
7. I felt that everything I did was an effort.	4.9	N	N	N	F	Y	Som
8. I felt hopeful about the future.	6.4	Y	N	N	F	N	Pos
9 I thought my life had been a failure.	0.8	N	N	N	F	N	Dep
10. I felt fearful.	1.3	N	N	N	F	N	Dep
11. My sleep was restless.	0.7	N	N	N	B	Y	Som
12. I was happy.	1.3	Y	N	N	F	N	Pos
13. I talked less than usual.	0.5	N	N	N	B	Y	Som
14. I felt lonely.	1.3	N	N	N	F	N	Dep
15. People were unfriendly.	9.1	N	Y	Y	F	N	Int
16. I enjoyed life.	1.3	Y	N	N	F	N	Pos
17. I had crying spells.	0	N	N	N	B	N	Dep
18. I felt sad.	0	N	N	N	F	N	Dep
19. I felt that people dislike me.	2.4	N	Y	Y	F	N	Int
20. I could not get "going."	0	N	N	N	B	Y	Som

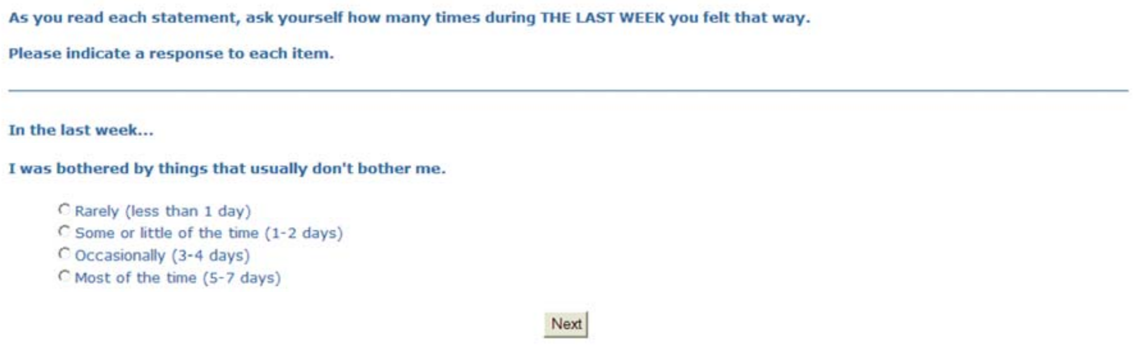
Note. Y = yes; N = no; Som = somatic symptoms; Dep = depressed affect; Pos = positive affect; Int = interpersonal problems.

Table 46. Results of the GEE Model for Testing the Effect of Identified Item Characteristics and Mean Difference Scores for Each Item Characteristic Across Assessments

				NCP		Same response	Different response	
	df	*Wald* χ^2	*p*	LL	UL	(%)	(%)	EM_{Diff} *(SE)*
Refers to others	1	9.84	.002	2.23	22.86			
Yes						81.3	18.7	.027 (.01)
No						80.5	19.5	.014 (.01)
Perception others	1	39.71	<.001	21.69	63.15			
Yes						86.1	13.9	.000 [a] (.02)
No						80.0	20.0	.041 (.01)
Behaviour/feeling	1	49.50	<.001	29.06	75.35			
Behaviour						82.0	18.0	-.009 [a] (.01)
Feeling						80.0	20.0	.050 (.01)
Reverse coded	1	23.86	<.001	10.50	42.63			
Yes						77.5	22.5	.003 [a] (.01)
No						81.4	18.6	.038 (.01)
Somatic Subscale	1	100.41	<.001	70.15	136.0			
Yes						78.0	22.0	.039 [a] (.01)
No						82.0	18.0	.002 (.01)

Note. Range of item responses 0-3. Theoretical range of difference scores -3 to 3. L=lower limit of 90% CI of non-centrality parameter; UL= upper limit of 90% CI of non-centrality parameter. Estimated marginal means are mean difference scores (first score-second). Positive numbers indicate higher item values on the first assessment compared to the second assessment.

[a] EM_{Diff} was statistically different in the model from the EM_{Diff} for the other category of the identified item characteristic.



As you read each statement, ask yourself how many times during THE LAST WEEK you felt that way.

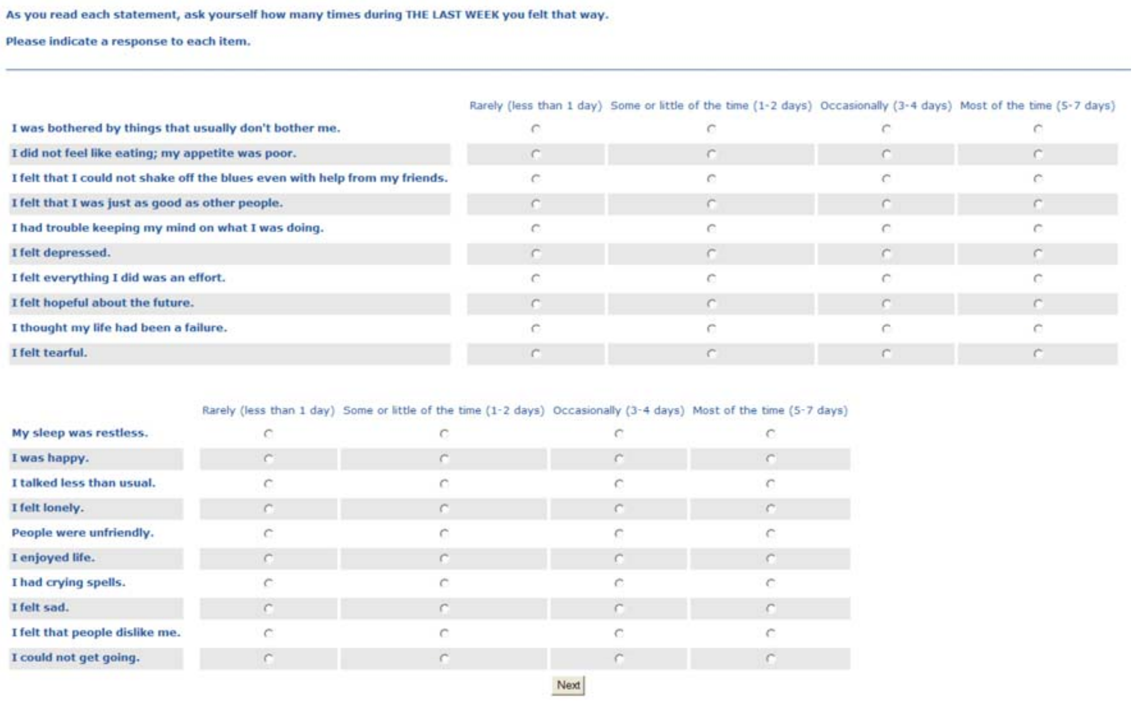
Please indicate a response to each item.

In the last week...

I was bothered by things that usually don't bother me.

○ Rarely (less than 1 day)
○ Some or little of the time (1-2 days)
○ Occasionally (3-4 days)
○ Most of the time (5-7 days)

Figure 1. **Example CES-D item in the one item per screen format.**



As you read each statement, ask yourself how many times during THE LAST WEEK you felt that way.

Please indicate a response to each item.

	Rarely (less than 1 day)	Some or little of the time (1-2 days)	Occasionally (3-4 days)	Most of the time (5-7 days)
I was bothered by things that usually don't bother me.	○	○	○	○
I did not feel like eating; my appetite was poor.	○	○	○	○
I felt that I could not shake off the blues even with help from my friends.	○	○	○	○
I felt that I was just as good as other people.	○	○	○	○
I had trouble keeping my mind on what I was doing.	○	○	○	○
I felt depressed.	○	○	○	○
I felt everything I did was an effort.	○	○	○	○
I felt hopeful about the future.	○	○	○	○
I thought my life had been a failure.	○	○	○	○
I felt tearful.	○	○	○	○

	Rarely (less than 1 day)	Some or little of the time (1-2 days)	Occasionally (3-4 days)	Most of the time (5-7 days)
My sleep was restless.	○	○	○	○
I was happy.	○	○	○	○
I talked less than usual.	○	○	○	○
I felt lonely.	○	○	○	○
People were unfriendly.	○	○	○	○
I enjoyed life.	○	○	○	○
I had crying spells.	○	○	○	○
I felt sad.	○	○	○	○
I felt that people dislike me.	○	○	○	○
I could not get going.	○	○	○	○

Figure 2. Example CES-D items in the multiple items per screen and response options to the right format.

Note. The font size of the items in Figures 1 and 2 appears different. When the items appear on the computer screen, the font size is the same.

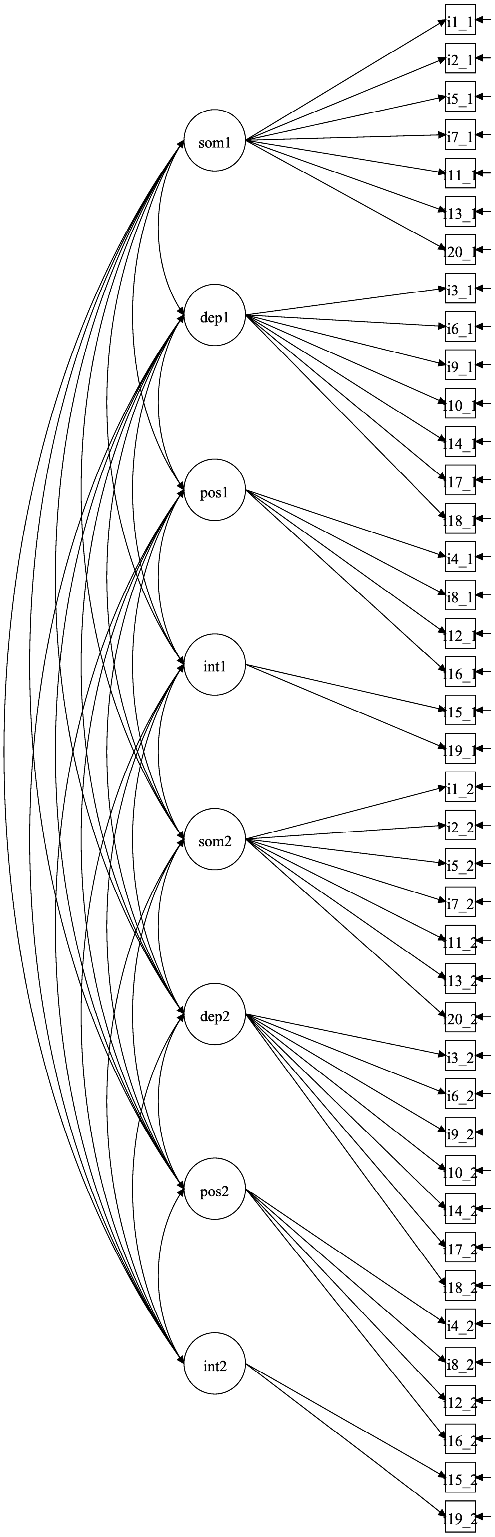


Figure 3. **Initial model structure for test of measurement invariance on first assessment compared to the second assessment.**

DV	S	E	T	F	S*T	S*F	E*T	E*F	T*F	S*E	S*E*T	S*E*F	S*T*F	E*T*F	S*E*T*F
Composite CES-D	•	•	•	•	•	•	•	•							
Somatic Symptoms	•	•	•	•	•	•	•	•							
Depressed Affect	•	•	•	•	•	•	•	•							
Positive Affect	•	•	•	•	•	•	•	•	•				•	•	
Interpersonal Problems	•	•	•	•	•	•	•	•	•					•	
CES-D item 1	•	•	•	•	•	•	•	•							
CES-D item 2	•	•	•	•	•	•	•	•							
CES-D item 3	•	•	•	•	•	•	•	•							
CES-D item 4	•	•	•	•	•	•	•	•	•				•	•	
CES-D item 5	•	•	•	•	•	•	•	•		•		•			
CES-D item 6	•	•	•	•	•	•	•	•							
CES-D item 7	•	•	•	•	•	•	•	•							
CES-D item 8	•	•	•	•	•	•	•	•							
CES-D item 9	•	•	•	•	•	•	•	•							
CES-D item 10	•	•	•	•	•	•	•	•							
CES-D item 11	•	•	•	•	•	•	•	•							
CES-D item 12	•	•	•	•	•	•	•	•	•				•		
CES-D item 13	•	•	•	•	•	•	•	•							
CES-D item 14	•	•	•	•	•	•	•	•		•		•			
CES-D item 15	•	•	•	•	•	•	•	•	•	•	•			•	
CES-D item 16	•	•	•	•	•	•	•	•							
CES-D item 17	•	•	•	•	•	•	•	•							
CES-D item 18	•	•	•	•	•	•	•	•							
CES-D item 19	•	•	•	•	•	•	•	•	•					•	
CES-D item 20	•	•	•	•	•	•	•	•							

Figure 4. Terms included in final models for LMM analysis for each dependent variable.

Note. S=sex; E=English fluency; T=time; F=format. White background denotes terms of interest for the study. Dots indicate term was included in final model.

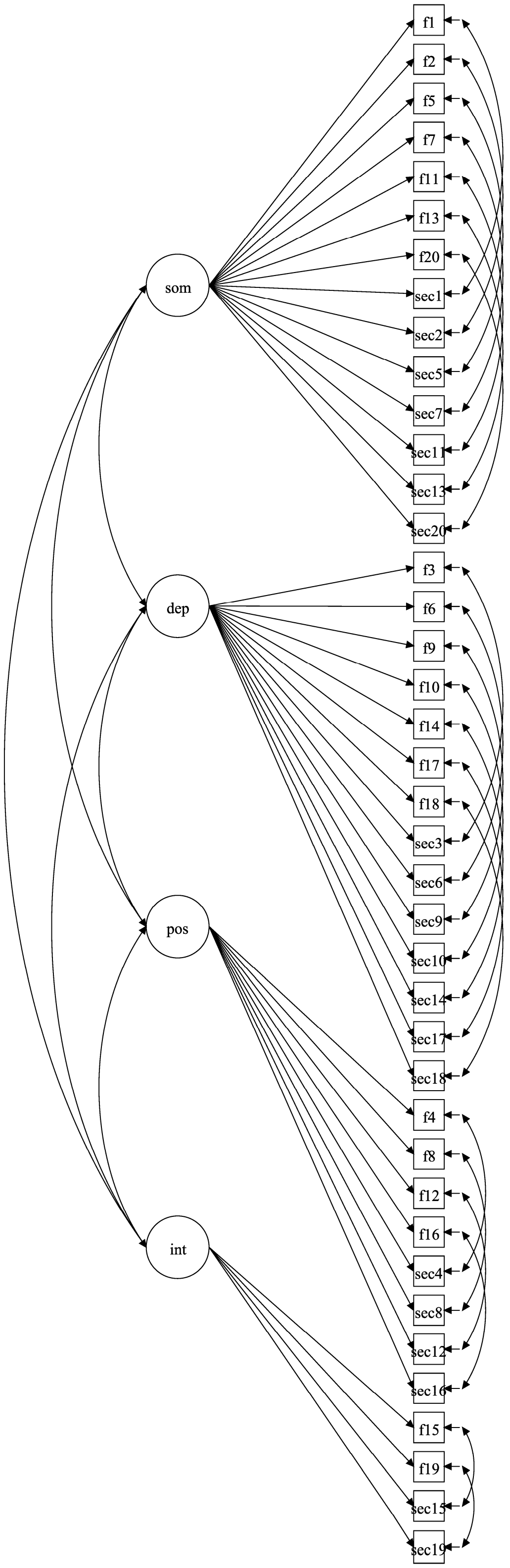
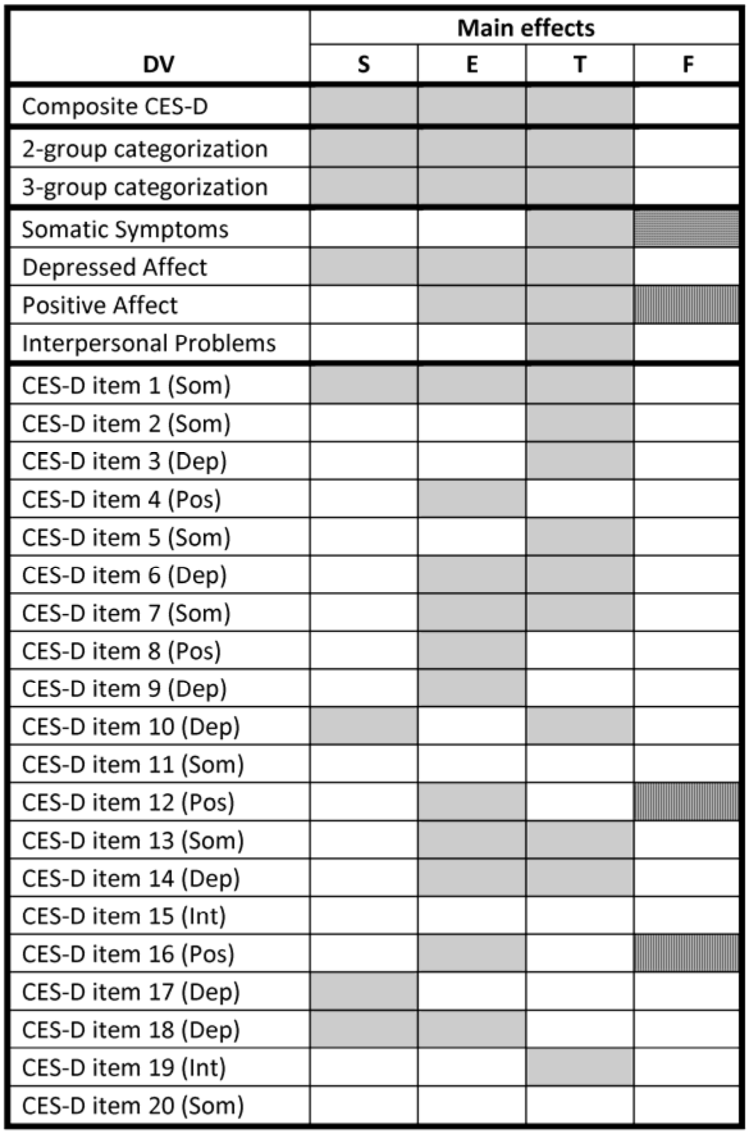


Figure 5. Structure of model used to evaluate measurement invariance on first assessment compared to the second assessment.

DV	Terms of Interest in the model								Additional terms with effects				
	S	E	T	F	S*T	S*F	E*T	E*F	T*F	S*E*T	S*E*F	S*T*F	E*T*F
Composite CES-D	░	░	░										
2-group categorization	░	░	░										
3-group categorization	░	░	░										
Somatic Symptoms			░	▤									
Depressed Affect	░	░	░										
Positive Affect		░	░	▥									
Interpersonal Problems			░										
CES-D item 1 (Som)	░	░	░	▤									
CES-D item 2 (Som)													
CES-D item 3 (Dep)			░										
CES-D item 4 (Pos)		░											
CES-D item 5 (Som)			░										
CES-D item 6 (Dep)		░	░				░						
CES-D item 7 (Som)		░	░										
CES-D item 8 (Pos)		░											
CES-D item 9 (Dep)		░											
CES-D item 10 (Dep)	░		░										
CES-D item 11 (Som)													
CES-D item 12 (Pos)		░		▥									
CES-D item 13 (Som)		░	░										
CES-D item 14 (Dep)		░	░								░		
CES-D item 15 (Int)									░				
CES-D item 16 (Pos)		░		▥									
CES-D item 17 (Dep)	░												
CES-D item 18 (Dep)	░	░											
CES-D item 19 (Int)			░										
CES-D item 20 (Som)													

Figure 6. Visual presentation of terms in the models of composite scores, categorizations, subscale scores, and items that were statistically significant when all terms of interest and appropriate higher order interactions are included in the models.

Note. S=sex; E=English fluency; T=time; F=format. Shaded terms were statistically significant, p<.05 for composite CES-D, p<.025 for 2 and 3-group categorizations, and p<.0125 for subscales and items. Effects for in each model for a given IV were in the same direction except Format, where effects in the same direction are noted with either vertical (multiple item per screen higher than one-item per screen) or horizontal lines (one item per screen higher than multiple items per screen).



DV	Main effects			
	S	E	T	F
Composite CES-D	■	■	■	
2-group categorization	■	■	■	
3-group categorization	■	■	■	
Somatic Symptoms			■	▤
Depressed Affect	■	■	■	
Positive Affect		■	■	▥
Interpersonal Problems			■	
CES-D item 1 (Som)	■	■	■	
CES-D item 2 (Som)			■	
CES-D item 3 (Dep)			■	
CES-D item 4 (Pos)		■		
CES-D item 5 (Som)			■	
CES-D item 6 (Dep)		■	■	
CES-D item 7 (Som)		■	■	
CES-D item 8 (Pos)		■		
CES-D item 9 (Dep)		■		
CES-D item 10 (Dep)	■		■	
CES-D item 11 (Som)				
CES-D item 12 (Pos)		■		▥
CES-D item 13 (Som)		■	■	
CES-D item 14 (Dep)		■	■	
CES-D item 15 (Int)				
CES-D item 16 (Pos)		■		▥
CES-D item 17 (Dep)	■			
CES-D item 18 (Dep)	■	■		
CES-D item 19 (Int)			■	
CES-D item 20 (Som)				

Figure 7. Visual presentation of statistically significant terms in main effects models of composite scores, categorizations, subscale scores, and mean item responses.

Note. S=sex; E=English fluency; T=time; F=format. Shaded terms were statistically significant, p<.05 for composite CES-D, p<.025 for 2 and 3-group categorizations, and p<.0125 for subscales and items. Effects for in each model for a given IV were in the same direction except Format, where effects in the same direction are noted with either vertical (multiple item per screen higher than one-item per screen) or horizontal lines (one item per screen higher than multiple items per screen.