Applying Computer Adaptive Testing Methods to Suicide Risk Screening in the Emergency Department

EDWIN D. BOUDREAUX PhD1, DEREK P. DE BEURS MSc, PhD2,3, TAM H. NGUYEN PhD, MSN/MPH, RN4, BRIANNA L. HASKINS MS5, CELINE LARKIN PhD2, BRUCE BARTON PhD6,

1Departments of Emergency Medicine, Psychiatry, and Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA
2Netherlands Institute for Health Services Research (NIVEL), Utrecht, The Netherlands
3Department of Clinical Psychology, Faculty of Psychology and Education, VU University Amsterdam, Amsterdam, The Netherlands
4Connell School of Nursing, Boston College, Chestnut Hill, MA, USA
5Department of Emergency Medicine, University of Massachusetts Medical School, Worcester, MA, USA
6Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA

We would like to acknowledge the project management of Donna Lesperance, MA, MPH. This study was funded through internal department funds. None of the authors have sources of potential bias or financial interests to disclose.

ABSTRACT
Objective: Combine test theory with technology to develop brief, reliable suicide risk measures in the emergency department.
Methods: A computer adaptive test for suicide risk was built using the Beck Scale for Suicide Ideation and tested among the emergency department population. Data were analyzed from a sample of 1,350 patients in several Massachusetts emergency departments. The test was built as outlined by the National Institutes of Health Patient-Reported Outcomes Measurement Information System.
Results: Of 1,350 patients, 74 (5%) scored above the cutoff of BSS > 2. Item 2, “Wish to die”, was the most informative item. When using only Item 2, 20% (n = 15/74) of at-risk patients and 3% (n = 40/1,276) of not-at-risk patients were misclassified. Patients were classified after four items with computer adaptive testing trait estimates highly comparable to those of the full scale. The precision rule model did not reduce the scale.
Conclusions: This study models the creation of a computer adaptive test for suicide ideation and marks the start of the development of computer adaptive
tests as a novel suicide risk screening tool in the emergency department.

Computer adaptive tests hold promise for revolutionizing behavioral health screening by addressing barriers including time and knowledge deficits.

The Joint Commission (2016) has identified the emergency department (ED) as a key venue for efforts to detect and reduce suicide risk. However, clinical realities usually prevent such screening from happening in a systematic way. Several aspects of screening impede successful implementation. First, suicide risk screening tools (Quinlivan et al., 2014) and individual risk factors (Larkin, Di Blasi, & Arensman, 2014) often have low specificity and tend to be overly inclusive in identifying potential suicide risk. Second, even when screening for a risk factor is performed, ED clinicians are rarely well-versed with how to further assess an individual who screens positive (Betz et al., 2013, 2015). These deficits are compounded by the inconvenience of using, scoring, and interpreting standardized paper-and-pencil measures that could help to quantify severity, such as the Beck Scale for Suicide Ideation (BSS; Beck & Steer, 1993). Finally, clinicians’ fidelity to screeners can be poor: Up to one-third of patients with negative suicide screens documented in their medical record stated that they had not been asked the suicide ideation or behavior items by their nurse (Boudreaux et al., 2016). Even the integration of a screener into an electronic health record does not guarantee complete fidelity.

Fortunately, computerized assessments sensitive to the “need for speed” and need for fidelity have been enhanced dramatically with the incorporation of computer adaptive testing (CAT) methods. CAT relies on item response theory (IRT; Linden, 2011), which uses the responses on individual items to estimate a participant's latent trait, that is, underlying construct of interest. IRT is used to derive several outcome parameters for each item in a measure. CAT can use these derived item parameters to map all items in a bank onto a common scale, which allows the adaptive administration of the most relevant items to each individual based on their prior response. As a result, the number of items administered in a questionnaire is markedly reduced without losing reliability or the power to discriminate problem severity (Meijer & Nering, 1999). Moreover, even if individuals are administered different items, their scores can be compared. A 2017 study by Gibbons et al. (2017) developed a CAT model for suicide risk assessment using a broad spectrum of items related to suicide, depression, and anxiety. This model was able to quickly and accurately measure risk with an average of 10 items, fewer items than necessary for nonadaptive screening methods.

This paper illustrates the steps used to build and test a CAT for suicide risk for the ED population using the BSS (Beck & Steer, 1993), a well-established measure to assess the level of suicidality, and a measure that is often used to classify a patient with high or low risk for future suicidal behavior. To simulate the steps, we analyzed data from participants who were screened for suicide risk using the BSS in several EDs in the Northeastern United States. This study follows similar methods used to develop a CAT of the BSS with a sample of Dutch psychiatric patients, which demonstrated that the 19 ideation items could be reduced to an average of four items without losing discriminative ability (De Beurs, de Vries, de Groot, de Keijser, & Kerkhof, 2014). Replicating this study in a US ED population provides a potential solution for recommendations from organizations like the National Action Alliance.
for Suicide Prevention, which have called for studying universal screening approaches for suicide risk (Boudreaux & Horowitz, 2014).

METHODS

Item Response Theory

Computer adaptive testing relies on IRT (Linden, 2011), which is based on the relationship between participants’ responses to individual items and their overall latent trait that the scale aims to assess. Suicide ideation is the latent trait examined in this study. IRT models can be used to derive a discrimination parameter, denoted \( a \), and \((n \text{ response categories} – 1)\) threshold parameters, denoted \( b_1 \) to \( b_{n-1} \), for each item in a measure. The discrimination parameter indicates how well an item relates to the total scale and can be compared to factor loadings. The discrimination parameter also describes how well an item can differentiate between individuals that have latent traits below and above the item location or difficulty level. The following guidelines can be used to interpret the discrimination parameter: low (0.01–0.64), moderate (0.65–1.34), high (1.35–1.69), and very high (>1.70; Baker, 2001); with discrimination parameters ≥1 generally agreed as acceptable (Hambleton, Swaminathan, & Rogers, 1991). The threshold parameters identify where on the continuum of the latent trait the item functions best; the higher the \( b \) parameter value, the more “difficult” the item (i.e., it functions at higher trait levels). The threshold parameters are also used to plot categorical response curves. Categorical response curves depict the probability of endorsing each category along a given trait continuum. Figure 1 illustrates an example of the categorical response curves for Item #2 (“Wish to die”) on the BSS. The three different curves (1—black, 2—red, and 3—green) represent the probability that a participant with a certain score on the latent trait, scaled along the x-axis, chooses “I have no wish to die,” “I have a weak wish to die,” or “I have a moderate to strong wish to die,” respectively. If a given participant has a latent trait of 0.1 (i.e., low level of suicide ideation), he/she has a 100% chance of selecting the first response option “I have no wish to die”; and the chances that he/she will select the second or third response option is approximately 0%. On the other hand, if a participant has a trait level of 4 (i.e., high level of suicide ideation), he/she has about 10% chance of endorsing the first response option (black curve), 25% chance of endorsing the second option (red curve), and a 65% chance of endorsing option 3 (green curve).

[FIGURE 1]

Computer Adaptive Testing

Within IRT, there is one item that provides the best mix of discrimination and threshold parameters and is valuable in CAT. Each participant must answer the same, most informative item. The CAT algorithm uses the IRT parameters to select the second most informative item, based on the response to the first item. New items are administered until the computer knows enough about the individual’s latent trait.

Selection of Items and Scale

The first step of developing a CAT is to identify an appropriate existing instrument with established reliability and validity (Beck & Steer, 1993; Boudreaux, Jaques, Brady, Matson, & Allen, 2015). In this study, we used the well-validated BSS which
uses 19 items to identify the presence and severity of suicide ideation and two
additional items on past suicide attempts. A 20-year follow-up study among 6,891
psychiatric outpatients found that a BSS total score of >2 increased the chance of
future suicidal behavior sevenfold (Brown, Beck, Steer, & Grisham, 2000).
Therefore, BSS >2 was used as the clinical threshold in the Dutch BSS study (De
Beurs et al., 2014). While the BSS only has 19 items, there are no commonly
accepted alternative suicide instruments with substantially more items. Furthermore,
it has already been successfully used in a published simulated CAT development
study (De Beurs et al., 2014).

**Data Collection**
The BSS was administered to 1,350 adult ED patients presenting to three EDs within
the University of Massachusetts Memorial Healthcare System. Part of the sample
\((n = 951)\) was collected in 2011 and is described in a published validation study
(Boudreaux et al., 2015). The remaining sample \((n = 399)\) was collected in 2013 and
has not yet been published. Data collection methods were identical in both studies
and are described in depth in Boudreaux et al. (2015).

All patients ≥18 years old were considered for participation. Individuals were
excluded if they could not engage in conversation (e.g., persistent altered mental
status, intractable pain or vomiting, intubation). The age, race, gender, and insurance
status of all participants were recorded. The studies were approved by the University
of Massachusetts Medical School Institutional Review Board. All participants signed
written informed consent and completed the assessment during their visit. The
assessment consisted of numerous items related to suicide risk; only BSS data are
used in this study. The BSS has a skip rule in which participants that score “0” on
both Items 4 and 5 (i.e., no active suicide ideation) do not answer the remaining 14
ideation items. As such, data for these 14 items were missing for most of the sample.

**Data Analytic Plan**
The data analytic plan followed generally accepted conventions as reported by the
National Institutes of Health Patient-Reported Outcomes Measurement Information
System (PROMIS) (Pilkonis et al., 2011) and de Beurs et al. (De Beurs et al., 2014).
To address the missing responses related to the “skip” pattern described above, we
used a strategy described in an article on measurement variability of the BSS (De
Beurs, Fokkema, de Groot, de Keijser, & Kerkhof, 2015) and assumed that all
participants that answered Items 4 and 5 as “0” would have scored “0” on the 14
other ideation items also. All analyses were performed for both the data set with
imputed “0s” and the original data set with the 14 items as missing, and item
parameters for both data sets will be reported. For all other statistics (e.g.,
assumptions, differential item functioning [DIF]), we will report results from the “0”
imputed data and only report if there was a difference with the original nonimputed
data set. Data were analyzed using the following steps as described by PROMIS
(Pilkonis et al., 2011).

*Calculate Descriptive Statistics*
Frequency, mean, and standard deviation were calculated for each item. Internal
consistency of the 19 BSS ideation items was calculated using Cronbach’s alpha. An
alpha of .80 was regarded as an acceptable minimum. Descriptive statistics were
calculated using the apply function in the R base package (Revelle, 2017). Cronbach’s alpha was calculated using the psych package (Revelle, 2017).

**Test Assumptions About the IRT Model**

The basic IRT assumptions of unidimensionality, local independence, and monotonicity (Baker, 2001; Linden, 2011; Pilkonis et al., 2011; Reeve et al., 2007) were tested. Unidimensionality implies that the scale measures only a single latent trait. Various fit indices, such as standardized root mean square residual of (SRMSR <0.08 for a good fit), can be used when confirming the one-dimensional structure via confirmatory factor analysis (CFA; Rosseel, 2012).

Local independence is related to unidimensionality and refers to the independence of items when the trait of interest is held constant. If we control for level of suicide ideation, the items should not correlate with each other. Local independence was analyzed by checking for scores >0.2 in the residual matrix of the CFA, a commonly used approach (Edelen & Reeve, 2007; Reeve et al., 2007).

Monotonicity is established if the chance of endorsing an item increases continuously as the underlying latent trait increases. Items will then have curves that display hierarchically (see Figure 1). Monotonicity will be checked by visually inspecting the order of the categorical response curves as well as by calculating a scalability coefficient for each item. Items that are not hierarchical and have a scalability coefficient of <.3 were flagged (van der Ark, 2007). Mokken scale analysis was performed using the mokken package (van der Ark, 2007). Within Mokken scale analysis, a rest score of each item is calculated by subtracting the response on an individual item with the sum score of all item responses. The ratio between the covariance of the item response and the rest score, and the maximum covariance gives the scalability coefficient. For more details, see van der Ark (2007).

**Fit the IRT Model to the Data**

Selecting an appropriate model depends on how item response categories are formatted and the overall model-data “fit.” For measures with items that have ordered response categories (i.e., Likert), like the ones on the BSS, the graded response model (GRM) is commonly used (Maydeu-Olivares, 2013; Samejima, 1969). The fit of this model to our data was tested using the M2 statistic (Cai & Hansen, 2013), a robust goodness-of-fit statistic developed for measures offering ordered response categories. The IRT parameters were fit using the ltm package (Rizopoulos, 2007).

**Evaluate Differential Item Functioning**

Differential item functioning (Smits, Cuijpers, & van Straten, 2011) exists if patient characteristics, such as age or gender, result in a different probability of endorsing an item, even with the same level of underlying trait. We evaluated DIF for gender and age (18–49 years, 50–69 years) using logistic regression. The magnitude of detectable DIF was evaluated by reporting the $R^2$ statistics (<.13 = negligible, between .13 and .26 = moderate or >.26 = large DIF). The Package Lordif was used to estimate DIF (Choi, 2016).
**Build the CAT Administration Algorithm**

We performed a post hoc CAT simulation using the IRT parameters from Step 3. In CAT, three steps are required: a starting point, a method for selecting the next item, and a stopping rule. The starting level of theta will be set at 0 (Smits et al., 2011). The first most informative item at this initial latent value is selected as the first item. All participants will start with this most informative item. After a participant answers the first item, the CAT model can estimate the latent trait based on the response selected, using Bayesian model Estimation with a poisson distribution as a prior estimate. Items were selected by maximizing fisher information at theta (Nydick, 2014; Smits et al., 2011). Finally, a stopping rule was selected. Selection depends on the underlying measurement goal. When the goal is to develop a screener to differentiate negligible versus non-negligible risk, a predetermined clinically important cutoff value can be used (classification rule). When the goal is to get a precise estimate of an individual’s trait score, the steps above can be applied to stop once an individual’s score is obtained within a specified standard error of measurement (precision rule). Three different stopping rules were tested for this study to illustrate both goals.

First, we manually classified patients into negligible risk and non-negligible risk based only on the scores of the most informative item as identified by the steps above. If a patient scored 0 on the most informative item, we classified them into negligible risk, and if patients scored >0 on this first item, we classified them as non-negligible risk. This classification is compared with the true total score of the patients on the BSS, where we used the often-used threshold of total score of BSS >2 as a cutoff. Next, to let the CAT classify patients automatically based on a predetermined cutoff, we used the theta that corresponded to an observed BSS score >2. For the current data, an observed BSS score of 2 is similar to an average theta score of 3.5. The CAT stops if it can classify a participant as having a latent trait of above or below this interval with a certainty of 95%. Finally, we simulated a precision rule CAT using different sizes of measurement error (ranging from 0.8 to 0.3). For each analysis, the CAT algorithm offers a vector of final CAT-based theta estimates. Pearson correlation was calculated to compare the final CAT-based theta estimates with the theta estimates given the entire item bank. The CAT was performed via the catIrt package (Nydick, 2014).

**RESULTS**

**Descriptive Statistics**

The average age of the sample was 44.6 (SD = 18), with 52% female, 88% White, 12% Latino, and 8% African American, which is very similar to the demographics of the EDs sampled (see Table 1). The Cronbach's alpha of the BSS was .95, which corresponds to high internal consistency. Of 1,350 participants, 1,185 (88%) had a total score of 91 (7%) scored 1 or 2, and 74 (5%) had a total BSS score of >2 (non-negligible risk).

**[TABLE 1]**

**IRT Model Assumptions**

A principal components analysis was selected in the exploratory phase to ensure a more reliable factor solution (Nunnally & Bernstein, 1994). It resulted in a one-factor
solution, with the first factor explaining 63% of the variance. The second factor explained 7% of the variance, yielding a ratio between the first factor and second factor of 9, which is much larger than the minimum ratio of 4. Confirmatory factor analysis also supported a one-factor solution (comparative fit index of 1.00, Tucker–Lewis index of 1.044, root mean square error of approximation of 0.0 standardized root mean square residual of 0.055). These results support the unidimensionality of the BSS. Local independence was confirmed by the residual correlation matrix showing no value > .2. Monotonicity was confirmed by no items having a scalability coefficient above .3. All categorical response curves were ordered hierarchically, except for Items 11, 13, and 19.

**IRT Model**

The IRT model for the “0” imputed data is presented in Table 2. The overall model fit was satisfactory (RMSEA 0.04, CI 90% 0.04–0.048, TLI = 0.99, CFI = 0.99). Given that each of the 19 ideation items on the BSS has three response categories, two location/difficulty parameters were estimated for each item, $b_1$ and $b_2$. Based on these, the BSS items largely function at the higher region of the latent suicide ideation scale, ranging from $b_1 = 2.214$ to 4.025 and $b_2 = 2.5$ to 4.597. Within this range, Items 1, 2, 4, 8, and 11 function at relative lower suicide ideation trait levels relative to Items 16, 17, and 18, which function at highest levels. The IRT model for the original (nonimputed) data is also presented in Table 2 for comparison. Due to the large amount of missing data, we were not able to test the GRM model fit for the original data.

| **TABLE 2** |
| **Differential Item Functioning** |
| To run the differential item function analysis, each item must have at least two valid response categories with five valid cases. From Table 1, it can be seen that DIF analysis was not possible for three items. None of the other items demonstrated statistically significant differential item function across gender and age groups. |

| **CAT Simulation** |
| Computer adaptive testing simulation shows that for all individuals, the starting item was Item 2 (“Wish to die”), because it offered the most discriminating information when the starting level of theta was set at 0. Table 3 shows the classification of patients if one would base their classification only on Item 2 (a patient has a non-negligible risk if Item 2 > 0). One thousand two hundred and fifty-one participants scored 0 on Item 2, of which 20% (15/74) would be misclassified (i.e., classified as having a negligible risk while their total BSS score would be > 2). Under the cutoff Item 2 > 0, 3% (40/1,276) would be classified as non-negligible risk while their total score was BSS ≤ 2. Of the 37 patients that score 2 on Item 2 (“Wish to die”), 6% (2/35) would have been classified as non-negligible risk, while their total score would be ≤ 2. In sum, based on the cutoff Item 2 > 0, 55 patients (4%) would be misclassified.

| **TABLE 3** |
| When the stopping rule was set at the predetermined cutoff of BSS > 2 in the CAT simulation, an average of 3.5 items (SD 2.1) was needed to classify a patient, an 82% |

This is a NIVEL certified Post Print, more info at [http://www.nivel.eu](http://www.nivel.eu)
reduction in items as compared to the original BSS ideation scale (3.5/19). The correlation between the observed scores on the full test and the trait (theta) scores using the CAT was 0.97, indicating a high level of similarity between the two estimates. When testing BSS using the precision stopping rule, the size of standard error was varied between 0.3 (strict) and 0.8 (lenient), but the mean number of items administered remained close to the original BSS.

**DISCUSSION**

Those who die by suicide often present to the ED in the year before death (Da Cruz et al., 2011; Morrison & Laing, 2011), and a significant proportion of nonpsychiatric ED patients endorse suicide ideation when screened (Allen et al., 2013; Claassen & Larkin, 2005). While universal suicide screening may be worthwhile, there are several implementation challenges associated with it. CAT holds tremendous promise for suicide screening because it can counter issues of clinician fidelity and interpretation of threshold scores.

Computer adaptive testings can be developed for two purposes, both of which are relevant to EDs: either to classify individuals into groups, such as for screening patients in or out of a risk group, or to precisely estimate a latent trait, such as suicide ideation severity. When using the most informative item as a screener, 15 (20%) of the patients at risk and 40 patients not at risk (3%) were misclassified. When we applied the CAT using a classification stopping rule based on a predetermined, clinically relevant, cutoff score, it could reduce the item administration for suicide ideation from 19 items to an average of 3.5, with an estimated theta that was highly correlated with the theta of the full scale. This is consistent with results published in the Dutch study (De Beurs et al., 2014). When we tested the BSS-CAT using a precision stopping rule, based on obtaining individual scores within a specified range of standard errors, the mean number of items administered did not significantly differ from the non-CAT format. This could be related to a lack of participants with a heightened level of suicidality within the sample. The sample was drawn from a general ED population, most of whom were not suicidal, making non-negligible suicide risk underrepresented. The BSS-CAT included a pool of items that largely functioned at “higher” suicide ideation trait levels, and most participants (n = 1,185; 88%) showed no suicide ideation. Therefore, there was not enough data for this BSS-CAT simulation to accurately estimate the underlying trait. The BSS-CAT needed at least 17 items before it could determine the trait level of a participant within a certain level of precision. If the underlying measurement goal is to precisely assess suicide ideation trait levels of individuals in this population, future research should include additional items at the lower and middle end of the suicide ideation trait scale, or data should be collected in a specific ED with more suicidal patients. This will likely lead to more accurate quantification of the severity of the problem while also reducing the average number of items needed to be administered.

The next step would be to validate the CAT against a criterion reference measure, such as a clinically administered suicide risk assessment or follow-up after the ED to detect prospective suicidal behavior. If validated, the CAT-based screener could improve the screening of suicide risk in the ED because only four items or less would be needed to classify a patient. Upon validation, the CAT would require further testing of acceptability and feasibility in the typical ED setting.
Although computer software is getting more user-friendly, running a CAT in real time still requires software that is not fully implemented in health care settings. The psychometrics center of the University of Cambridge offers an open-source platform (https://www.psychometrics.cam.ac.uk/newconcerto). The PROMIS initiative launched an online assessment center, where users may build a CAT if they have access to the IRT parameters (https://www.assessmentcenter.net/). They also have a collection of scales for which they already built a CAT, which may be used for research purposes. For more details on available online platforms, refer to Oppl (2017). As the field is still rapidly developing, novel algorithms and software will continue to emerge.

LIMITATIONS

The distribution of the BSS total scores is highly skewed, with most participants scoring 0 ($n = 1,185$ (88%). As the individual item category response curves looked satisfactory, we do not expect this to influence the IRT parameters. However, its effect on the CAT simulation is not clear. Therefore, the results of this article should be regarded as a demonstration of CAT, rather than a definite result. The imputation of zeros for missing items that were not administered because of the instructions on the BSS (i.e., imputed data) is problematic and demonstrates some of the obstacles that may arise when using existing data or instruments with branching logic as the starting point of a CAT. Moreover, because of the missing values, it was not possible to run an accurate CAT on the original, nonimputed data. To improve the accuracy of the item parameters, future studies should ask respondents to answer all items of the BSS, whether or not they have suicide ideation. Furthermore, it is likely that at least some patients were reluctant to report suicidal thoughts due to shame or fear of being stigmatized. Indeed, many studies indicate that patients with suicidal behavior are being stigmatized and treated with a lack of empathy (Artis & Smith, 2013; Pompili, Girardi, Ruberto, Kotzalidis, & Tatarelli, 2005), resulting in an underestimation of patients with suicidal thoughts at EDs. An associated limitation is that our item bank was relatively small having only used items in the BSS. A more comprehensive CAT item bank including nonsuicide-related items could help to overcome patient self-stigma or reticence in disclosing suicide ideation. Concealment of suicidality can occur to avoid potential negative consequences (Denmark, Hess, & Becker, 2012) Nonsuicide-related items (i.e., sleep, irritability, and impulsivity) that tap ideation trait levels could indicate risk without requiring the patient to explicitly endorse a suicide-related item. The CAT suicide risk assessment tool is only feasible when computerized data collection is possible. Due to the financial and human resources often involved with the implementation of computerized data collection, application is limited. Barriers such as the cost of equipment, software, training, and staffing could limit the feasibility of CAT implementation. Finally, the distribution of the BSS total scores is highly skewed, with most participants scoring 0 ($n = 1,185$ (88%). As the individual item category response curves looked satisfactory, we do not expect this to influence the IRT parameters. However, its effect on the CAT simulation is not clear. Therefore, the results of this article should be regarded as a demonstration of CAT, rather than a definite result.
CONCLUSIONS
This study demonstrated the creation of a CAT for suicide ideation and marks a starting point for the development and implementation of CAT as a novel suicide risk screening tool in the ED. Such a tool could help to overcome existing issues of time constraints and suboptimal fidelity in suicide screening in the ED. Our analyses also highlighted the challenges associated with building CATs based on different underlying measurement goals (i.e., to “categorize” individuals, or to get a “precise” estimate of a trait score). An item bank may work well for one purpose but not the other. Additional research may be needed to further develop the CAT, validate it against an external criterion, and to test its implementation in a real-world setting.

REFERENCES


This is a NIVEL certified Post Print, more info at http://www.nivel.eu
FIGURES AND TABLES

Figure 1: Category response curve of Item 2. Black curve: probability of answering “I have no wish to die,” Red: “I have a weak wish to die,” and Green: “I have a strong wish to die.” X: level of suicidality, Y: probability of a response.

Item 2: Wish to die
Table 1. Descriptive Statistics of the Single Items of the BSS Using the Imputed Zero's Data

<table>
<thead>
<tr>
<th>Single-item content</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Wish to live</td>
<td>0 1 2</td>
</tr>
<tr>
<td>2. Wish to die</td>
<td>0 1 2</td>
</tr>
<tr>
<td>3. Reasons living/dying</td>
<td>0 1 2</td>
</tr>
<tr>
<td>4. Desire to kill oneself</td>
<td>0 1 2</td>
</tr>
<tr>
<td>5. Save my life</td>
<td>0 1 2</td>
</tr>
<tr>
<td>6. Periods thinking about suicide</td>
<td>0 1 2</td>
</tr>
<tr>
<td>7. Frequency of thinking about suicide</td>
<td>0 1 2</td>
</tr>
<tr>
<td>8. Acceptance of idea of suicide</td>
<td>0 1 2</td>
</tr>
<tr>
<td>9. Ability to not commit suicide</td>
<td>0 1 2</td>
</tr>
<tr>
<td>10. Reasons for not committing suicide</td>
<td>0 1 2</td>
</tr>
<tr>
<td>11. Reasons for wanting to commit suicide</td>
<td>0 1 2</td>
</tr>
<tr>
<td>12. Specific plan to commit suicide</td>
<td>0 1 2</td>
</tr>
<tr>
<td>13. Access to suicide method</td>
<td>0 1 2</td>
</tr>
<tr>
<td>14. Courage/ability to commit suicide</td>
<td>0 1 2</td>
</tr>
<tr>
<td>15. Expectation to commit suicide</td>
<td>0 1 2</td>
</tr>
<tr>
<td>16. Preparations for suicide</td>
<td>0 1 2</td>
</tr>
<tr>
<td>17. Writing of suicide note</td>
<td>0 1 2</td>
</tr>
<tr>
<td>18. Arrangements for after suicide</td>
<td>0 1 2</td>
</tr>
<tr>
<td>19. Conceal ideation</td>
<td>0 1 2</td>
</tr>
</tbody>
</table>

BSS, Beck Scale for Suicide Ideation.

Table 2. IRT Parameters for “0” Imputed for Missing Ideation Data and Original Data

<table>
<thead>
<tr>
<th>Item and content</th>
<th>Parameters for “0” imputed for missing ideation data</th>
<th>Parameters for original data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Wish to live</td>
<td>Α 1.611 β1 2.475 β2 3.622</td>
<td>Α 2.974 β1 1.934 β2 2.929</td>
</tr>
<tr>
<td>2. Wish to die</td>
<td>Α 1.359 β1 2.298 β2 3.679</td>
<td>Α 3.666 β1 1.665 β2 2.362</td>
</tr>
<tr>
<td>3. Reasons living/dying</td>
<td>Α 1.388 β1 2.518 β2 3.818</td>
<td>Α 2.581 β1 1.874 β2 2.689</td>
</tr>
<tr>
<td>4. Desire to kill oneself</td>
<td>Α 2.609 β1 2.379 β2 3.321</td>
<td>Α 3.478 β1 2.148 β2 2.876</td>
</tr>
<tr>
<td>5. Save my life</td>
<td>Α 1.947 β1 2.567 β2 4.059</td>
<td>Α 2.984 β1 2.318 β2 2.937</td>
</tr>
<tr>
<td>6. Periods thinking about suicide</td>
<td>Α 2.562 β1 2.884 β2 3.537</td>
<td>Α 3.270 β1 2.103 β2 2.745</td>
</tr>
<tr>
<td>7. Frequency of thinking about suicide</td>
<td>Α 2.680 β1 2.817 β2 3.823</td>
<td>Α 2.973 β1 2.169 β2 3.084</td>
</tr>
<tr>
<td>8. Acceptance of idea of suicide</td>
<td>Α 3.115 β1 2.386 β2 3.635</td>
<td>Α 3.231 β1 1.697 β2 2.990</td>
</tr>
<tr>
<td>9. Ability to not commit suicide</td>
<td>Α 2.745 β1 3.042 β2 4.249</td>
<td>Α 2.980 β1 2.320 β2 3.451</td>
</tr>
<tr>
<td>10. Reasons for not committing suicide</td>
<td>Α 2.687 β1 2.626 β2 4.195</td>
<td>Α 2.427 β1 1.955 β2 3.046</td>
</tr>
<tr>
<td>11. Reasons for wanting to commit suicide</td>
<td>Α 2.769 β1 2.214 β2 2.540</td>
<td>Α 0.598 β1 −2.832 β2 0.340</td>
</tr>
</tbody>
</table>

This is a NIVEL certified Post Print, more info at [http://www.nivel.eu](http://www.nivel.eu)
### Table 3. Relating the Response on Item 2 to the Final Classification Based on the Total BSS

<table>
<thead>
<tr>
<th>Response on Item 2</th>
<th>Negligible risk (total BSS ≤ 2)</th>
<th>Non-negligible risk (total BSS &gt; 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.236</td>
<td>15</td>
</tr>
<tr>
<td>1</td>
<td>38</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>35</td>
</tr>
</tbody>
</table>