Published in final edited form as:

Multivariate Behav Res. 2019 ; 54(2): 159–172. doi:10.1080/00273171.2018.1503075.

Gleaning Information for Cognitive Operations from Don't Know Responses in Cognitive and Non-Cognitive Assessments

Edward H. Ip^{1,2}, Michelle F. Magee^{3,4}, Gretchen A. Youssef³, and Shyh-Huei Chen¹

¹Department of Biostatistical Sciences, Wake Forest University School of Medicine

²Department of Social Sciences & Health Policy, Wake Forest University School of Medicine

³MedStar Diabetes Institute

⁴Georgetown University School of Medicine

How the response category of Don't Know (DK), broadly defined, should be analyzed is a challenging issue in educational and psychological measurements. In educational tests that feature multiple-choice items, DK includes omitted responses (none of the items checked), and the option "Don't Know the Answer" if the option is available as part of the response categories. In low-stakes cognitive tests, DK can be used as a tool to not force test takers to guess; knowing that the test taker does not know the answer can lead to better understanding of deficits in ability and the related cognitive process.

A prototypical example of an explicit DK response category in a low-stakes cognitive test is assessment of knowledge about health – which is important, for example, for understanding compliance and self-management of chronic disease. An example of such a cognitive test is analyzed later in our real data example. The test, the "KNOW Diabetes" diabetes knowledge Test (DKT), consists of 13 items plus an additional two that are required only if a respondent requires insulin treatment for the disease. A sample question is: "High blood sugar may make you: A. Feel shaky and nervous, B. Feel full of energy, C. Have increased urination, thirst and hunger, D. Do not know." The response is coded as correct (key is C), not correct, or DK. It is desirable to know if a respondent actually does not know the answer. Without the DK option, a respondent who does not have the knowledge would be forced to guess or omit the item which increases the likelihood that a knowledge deficit which should be addressed will not be identified during the survey process.

Another example drawn from educational testing is when the test is given to a diverse population, such as in large-scale international assessments. Some of the students, even of the same grade, may not have learned a topic as part of their local curriculum while others may have. In this situation, allowing the student to say DK to a specific item is reasonable as the student may not have been previously exposed to curriculum materials required to solve the question at hand. DK is more often explicitly included in non-cognitive tests such as social and attitudinal surveys. Some variants of the DK option are "I don't care," "No opinion," or "Neutral" although one could argue that these responses are different from not knowing the answer.

The inclusion of DK in non-cognitive tests is a controversial issue, and researchers have argued both for and against such practice (Bishop, 2005; Krosnick et al., 2002). Reasons for inclusion include improving reliability and allowing those who do not know or have an opinion to express their view directly. Reasons for exclusion include concerns about encouraging respondents to take an “easy” path and confounding respondents so that they do not understand the question (Feick, 1989). For factual items, it seems reasonable to include DK as an option.

To distinguish between DK and the other response categories, we use the term “directional responses” to refer to the set of responses that indicate either a positive or negative direction. Example of direction is Agree/Disagree in a non-cognitive test or survey, and Correct/Incorrect in a cognitive test. Some existing methods preprocess non-directional categories, including DK, and treat them as directional. For example, in a cognitive test, DK (non-directional) is coded Incorrect (directional).

This paper discusses the application of an extension of the linear logistic test model (LLTM; Fischer, 1973), to both cognitive and non-cognitive tests that contain the DK option, in addition to the usual directional response categories. The analysis of multiple responses with DK options for the purpose of inferring cognitive operations underlying a construct of interest – ability, attitude, or belief – is a psychometrically challenging issue. In our diabetes test example, item responses, including possible DK responses, were collected from the same individual and therefore are not independent. Thus, traditional regression such as logistic regression cannot be directly applied to the observations. In item response theory (IRT; Lord, 1980), a commonly used method for scaling multiple responses, the response a person gives to a specific item can be explained by both a latent person effect (often called “ability”) and an effect due to the item itself. Successful applications of IRT include cognitive tests such as in large-scale educational assessments, and in recent years, non-cognitive tests such as personality assessment, attitude measurement, health assessment, and psychopathological assessment (Reise & Waller, 2009).

As an extension of IRT, the LLTM captures both the person effect and effects arising from various cognitive operations required for positive responses to a stimulus – e.g., correctly answering a math question. The LLTM is a generalization of the Rasch model (Rasch, 1960) in which the responses are explained both in terms of a person effect and the effects of item attributes, which are modeled as linearly additive components as a substitution for the so-called difficulty parameter in an IRT model. In other words, the LLTM is a parametric model specified to a formalized theory regarding the cognitive components required to solve items in a test. Embretson (1995), for example, identified seven cognitive operations on three knowledge types (factual/linguistic, schematic, and strategic) required for a correct response to mathematical word problems. The formal model for knowledge was based on a theoretical model of Mayer, Larkin, and Kadane (1984). A Q-matrix (of dimension = number of items x number of operations [Tatsuoka, 1983]) is prespecified to define the required cognitive operations for individual items. An entry of 1 in the matrix represents a required operation, and 0 otherwise. Partly because of its flexibility, the LLTM serves as the workhorse for many cognitive diagnosis models (e.g., see Dimitrov & Raykov, 2003). Subjects of

applications included math (Chen, MacDonald and Leu, 2011) and reading (Embretson and Wetzel, 1987).

While the LLTM was originally designed to delineate the cognitive difficulty of an item in terms of concrete item features – thus making it possible to design a test with a large number (in theory an infinite number) of items at any level of difficulty (e.g., Gorin, 2005) – the model has also found numerous other applications (Kubinger, 2008). These applications include the analysis of the Latin Square Task (Zeuch, Holling, & Kuhn, 2011), modeling of anger in psychological experiments (Ip, Smits, & De Boeck, 2009), detection of position effects (Hohensinn, Kubinger, Reif, Holocher-Ertl, Hkorramdel, & Frebort, 2008), study of teachers' ratings (Draney & Wilson, 2008), and fatigue and speeded item presentation effects (Kubinger, 2009). However, as far as we know, no LLTM has been developed for directly modeling DK responses.

The current research builds upon the work of Ip, Chen, and Quandt (2016), which proposes an IRT-based model for handling DK options in a test that consists of multiple items. Their proposed IRT model forms an integrated framework within which the estimated ability can be inferred from the directional responses as well as the DK responses. We extend that work to incorporate an explanatory model using item attributes for coding cognitive operations. In other words, information that is contained in the DK response contributes to the inference of a person's ability (in cognitive tests) or a latent construct of interest (in non-cognitive tests), as well as the required cognitive operations.

The work presented offers several advances to the measurement literature. First, it describes a novel LLTM that can handle directional responses in conjunction with DK responses under an integrated explanatory IRT. Second, the current work provides a method that could make use of existing software for estimating ability, as well as the LLTM parameters for a test that contains DK options. Finally, it demonstrates how the method can be practically applied to both cognitive and non-cognitive tests with DK options.

The article presents the following: First, we provide the background and method for the LLTM with DK option. The theoretical framework for analyzing DK responses is then introduced, and a method that allows the use of existing software programs to estimate both LLTM item parameters and latent ability is described. The next section presents two small simulation experiments for showing parameter recovery. The Real Data Analysis section presents a non-cognitive data set and a cognitive knowledge test for illustrating the approach. Finally, a discussion is included.

Method

The LLTM extends the Rasch model (Rasch, 1960) by allowing the kernel of the response function (logit of probability of positive response given ability) to be decomposed into the sum of person trait and a compensatory term of some weighted linear function of item attributes. In a test of I items and J examinees, for each item, the item characteristic curve (ICC) from the Rasch model is given by:

$$P(Y_{ij} = 1 | \theta_j) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \quad (1)$$

where Y_{ij} denotes the item response of the j^{th} person to the i^{th} item, θ_j denotes the (latent) trait score of the person j , and b_i is the item-difficulty parameters of item i , $i=1, \dots, I$. In the LLTM, the item difficulty parameter b_i is expressed as a linear combination of cognitive operations (Fischer, 1973, 1995):

$$b_i = \beta_0 + \sum_{k=1}^K q_{ik} \beta_k \quad (2)$$

where β_k denotes the impact of the k th cognitive operation; q_{ik} is an entry in the so-called Q-matrix, which is a predetermined weight assigned to cognitive operation k ; and β_0 is an intercept term common to all items. For calibration of the model, the LLTM requires the estimation of the parameters $\beta = (\beta_0, \beta_1, \dots, \beta_K)$. From a regression point of view, the cognitive operations are treated as fixed effects which are parameterized by fixed weights (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). Rijmen and De Boeck (2002) discussed random weights LLTM. In this paper we examine only fixed weights LLTM.

The inclusion of the DK category into the LLTM can be operationalized through the partially ordered set (poset) theory. The response categories – Correct, Incorrect, and DK – are conceptualized to form a poset structure, which is depicted in Figure 1(a). It can be proved that any poset structure can be hierarchically decomposed into ordinal and nominal structures such that individual models can be applied to the reparameterized categories (Theorem 1, Zhang & Ip, 2012). The decomposition theorem can be applied to general poset, but here we focus only on the structure in Figure 1(a).

The hierarchical decomposition results in two models: Model A: Condition of the response being directional, the two responses Correct (C) and Incorrect (I) forming two distinct categories; and Model B: the directional categories (C and I together) and the DK category form two distinct (non-directional) categories. Mathematically, the two item response models can be expressed as follows: Model A:

$$P(Y_{ij} = C | \theta_j, Y_{ij} = \{C, I\}) = \frac{\exp[\theta_j - (\beta_0 + \sum_{k=1}^K q_{ik} \beta_k)]}{1 + \exp[\theta_j - (\beta_0 + \sum_{k=1}^K q_{ik} \beta_k)]} \quad (3)$$

and Model B:

$$P(Y_{ij} = \{C, I\} | \theta_j) = \frac{\exp[\theta_j - (\beta_0^* + \sum_{k=1}^K q_{ik} \beta_k)]}{1 + \exp[\theta_j - (\beta_0^* + \sum_{k=1}^K q_{ik} \beta_k)]} \quad (4)$$

where β_0^* can be written as a shift term from the intercept term β_0 in Model A and can be expressed as follows:

$$\beta_0^* = \beta_0 + \delta, \quad (5)$$

Where δ is the shift parameter. Equation (5) uses the DK response as a reference category (coded 0). The actual reference category needs to be determined; see the explanation later in this section. Equations (3) - (5) together specify the LLTM with DK option. Model A, as specified in Equation (3), is the usual LLTM with Correct/Incorrect response categories whereas Model B is the additional component that models the DK option. Model B, in this sense, tries to glean additional information from the DK response for inferring latent ability and item attributes. The approach circumvents the need to make assumptions such as that DK is no different from Incorrect. As argued before, DK, assuming that it is indeed a valid response, itself contains information about the subject although this information may be relatively weak compared to the signal in the Correct/Incorrect dichotomy. One assumption that is implicit in Equation (3)-(5) is that given latent ability, the regression coefficients for cognitive operations across Model A and Model B are uniform except for the shift in the intercept.

The technical challenge for simultaneously estimating both Models A and B is that there are several elements – namely, the individual's latent ability (θ_j , a random variable) and the coefficients for cognitive operations β_k , $k = 1, \dots, K$ – common to both models. Separately estimating Models A and B is possible, but that would result in two sets of ability estimates for test takers as well as two sets of estimates for cognitive operations. The challenge is non-trivial, and Ip et al. (2016) proposed a recoding scheme to resolve this issue in IRT (not in LLTM). The idea is that for each item, Models A and B can be reformulated such that the responses can be treated as coming from two distinct subitems instead of from one single item. Intuitively, one can think of the DK option as providing a second piece of information for the latent ability above and beyond that provided by the directional Correct/Incorrect response. Specifically, Table 1 shows how an item with the DK option can be recoded into two distinct subitems.

The general result of equivalence between the poset and the subitem approach – i.e., the validity of the recoding scheme in IRT – was proved in Ip et al. (2016). We shall not repeat the proof here but would like to point out that (1) it is easy to prove that the same recoding scheme works for LLTM and (2) the logic of the general proof for poset LLTM easily follows the logic in the proof for IRT.

An important implication of the recoding scheme is that one can estimate the LLTM parameters and the latent abilities for items with DK simultaneously without developing new codes. Existing LLTM/IRT software programs can be used for model calibration and ability estimation. In other words, the parameter set $\beta_0^*, \beta_1, \dots, \beta_K$ and the latent abilities θ_j , $j=1, \dots, J$, can all be estimated using standard software. For the purpose of this paper, we used SAS PROC GLIMMIX. The program GLIMMIX is a highly flexible, efficient, and robust tool for

mixed effect modeling. As De Boeck and Wilson (2004) showed, the IRT and LLTM are generalized mixed models; thus, generic programs for fixed- and random-effects regression models such as GLIMMIX or NLMIXED can be readily deployed. For the LLTM considered here, it can be seen that once the DK option is recoded as subitem, generic programs also readily apply because LLTM is essentially a random-effect regression model. Our experience showed that for generic statistical programs such as GLIMMIX, estimating an LLTM for a medium-sized data set ($n=500$ and number of items=30) required only a few seconds. For ordinal data with DK options, or other more complex poset response structures, it is possible to use PROC NLMIXED for calibration. See De Boeck and Wilson (2004) for examples of NLMIXED.

Although in theory the latent ability can be estimated simultaneously in Models A and B using the recoding scheme, the question of how to code DK remains. The question is, if the knowledge about the DK response can be used to inform latent ability, which is directional, then should DK be coded as 0 or 1 in Model B? For example, if DK tends to contribute positive information about the examinee as compared to the directional categories (Correct/Incorrect) – i.e., higher ability tends to select DK compared to the other two categories combined – then DK should be coded as 1. Unfortunately, in practice, this is often not obvious to determine a priori if DK contributes to positive or negative information when compared to the other two categories. We proposed a data-driven solution: preprocess the data using a model with a slope (discrimination) parameter such that the estimated discrimination parameter is allowed to be either positive or negative. Items with negative parameters can then be reverse coded so that all discrimination parameters would eventually satisfy the positivity constraint.

Specifically, to determine the coding of the DK category for each item, we first calibrate the test, treating DK as missing, and then score the individuals. By using the individuals' scores, we then fit the following 2-parameter (2PL) logistic version of Model B with DK being coded 0 and the other two categories {C,I} coded 1:

$$P(Y_{ij} = \{C, I\} | \theta_j) = \frac{\exp(a_i \theta_j - b_i)}{1 + \exp(a_i \theta_j - b_i)}, \quad (6)$$

where a_i denotes the discrimination parameter, and θ_j is fixed at the estimated latent ability derived from the scoring procedure in the first step. If the fitted 2PL model indicates that $a_i < 0$, then for the specific item i , DK would be recoded as 1. Otherwise, DK remains coded as 0. Subsequently, this coding scheme for DK is applied to LLTM Model B as described above.

In contrast to alternative approaches such as placing DK between correct and incorrect so the response categories are still rank ordered, the current approach does not require such ad hoc decisions. The advantage of the data-driven approach is even clearer when one considers multiple response categories such as in Figure 1(b). Categories Agree, Neutral, and Disagree apparently form a logical rank-ordered list. To locate non-directional categories such as DK and Refuse to Respond within the directional list would be challenging unless the researcher

has a priori knowledge about the response mechanism. Another advantage of the proposed approach is that the information DK contributes to the underlying trait is allowed to vary across items, both in terms of direction (as determined using the 2PL model in the first step of preprocessing) and magnitude.

The models in Equations (3) and (4) can be further extended to include person covariates:

$$P(Y_{ij} = C | \theta_j, Y_{ij} = \{C, I\}) = \frac{\exp[\theta_j + \sum_{r=1}^R \gamma_r X_{jr} - (\beta_0 + \sum_{k=1}^K q_{ik} \beta_k)]}{1 + \exp[\theta_j + \sum_{r=1}^R \gamma_r X_{jr} - (\beta_0 + \sum_{k=1}^K q_{ik} \beta_k)]}, \quad (7)$$

and

$$P(Y_{ij} = \{C, I\} | \theta_j) = \frac{\exp[\theta_j + \sum_{r=1}^R \gamma_r X_{jr} - (\beta_0^* + \sum_{k=1}^K q_{ik} \beta_k)]}{1 + \exp[\theta_j + \sum_{r=1}^R \gamma_r X_{jr} - (\beta_0^* + \sum_{k=1}^K q_{ik} \beta_k)]}, \quad (8)$$

where γ_r is the regression coefficient for person covariates and X_{jr} is the person covariate. The implementation of the LLTM with person covariates is a trivial extension of the generalized linear mixed model using PROC GLIMMIX. It should be noted that such an extension, while only representing small efforts in terms of analysis and computation, could be critical for applications in which personal predictors are of interest in LLTM. For example, in an intervention study designed to evaluate a new method of classroom instruction, the intervention status of the individual can be readily included in the set of person predictors. We shall see an illustration of person covariates in the Real Data Analysis section.

Simulation Experiments

Simulation Study 1 (SS1): Parameter Recovery without DK Response

The purpose of SS1 is to examine if the regression parameters in the basic LLTM model in (3), i.e., one *without* DK response, can be accurately recovered using existing generalized mixed model software. We adopted the commonly used mixed model program SAS PROC GLIMMIX for the simulation experiments and real data analysis. The result from SS1 would provide a reference for the second simulation experiment, in which DK responses were included in the model such that the performance of the LLTM procedure for the two models can be directly compared.

SS1 used two Q-matrices, each with five attributes and 18 items. In the first Q-matrix, all attribute coding was distinct, i.e., the Q-matrix is full rank, whereas in the second, some items had identical vectors of attributes. The items and Q-matrices together were labeled Set A and Set B. Item responses were generated using Equation (3) with $K=5$. Two levels of number of persons were used: $N=200$ and $N=1,000$ for each set. We also combined Sets A and B so a new 36-item set was formed, and $N=1,000$ was used. Therefore, there were a total of $2 \times 2 + 1 = 5$ experimental conditions. For each condition, $M=100$ replications were

used to evaluate parameter recovery. Two criteria, bias and root mean squared error (RMSE), were used for the evaluation. Specifically, for the parameter $\beta_k, k = 1, \dots, 6$, bias and RMSE are defined as:

$$\text{Bias} = \frac{\sum_{m=1}^M (\hat{\beta}_{mk} - \beta_k)}{M}, \quad (9)$$

and

$$\text{RMSE} = \sqrt{\frac{\sum_{m=1}^M (\hat{\beta}_{mk} - \beta_k)^2}{M}}, \quad (10)$$

where $\hat{\beta}_{mk}$ represents the estimate for the parameter β_k at the m th replication.

Table 2 shows the result of SS1. The result shows that parameters in the LLTM can be recovered satisfactorily. The average biases and RMSEs are all close to zero for each of the parameters of interest.

Simulation Study 2 (SS2): Parameter Recovery with DK Response

The same item parameter and Q-matrices from Sets A and B from SS1 were used in SS2. Additionally, we included items that contain DK responses. Three items that contained DK responses were added to each of set A and set B, and the sets were respectively labeled DKA and DKB. Item-specific shift parameters were labeled $\delta_1, \dots, \delta_6$. The structure of the generative model was similar to the non-cognitive real data set. The responses to the items with DK responses were generated using Equation (4) with $K=5$, and their LLTM parameters were labeled $(\beta_0^*, \beta_1, \dots, \beta_5)$. Actual response data were generated by the application of Eq. (4) and, conditional on the condition that no DK response has been generated, followed by the application of Eq. (3)).

We created five levels of test structure: (1) Set A + DKA, (2) Set B + DKB, (3) Set A + Set B + DKA, (4) Set A + Set B + DKB, and (5) Set A + Set B + DKA + DKB. For each test structure condition, we used two levels of number of persons, $N=200$ and $N=1,000$.

Therefore, there were altogether a total of $2 \times 5 = 10$ conditions. For each condition, 100 replications were used for the evaluation of parameter recovery.

Table 3 shows the results from SS2. The biases and RMSEs for LLTM Model A (Equation (3)) are quite similar to those in SS1. Additionally, the biases and RMSEs for the LLTM Model B (Equation (4)) are also close to zero, suggesting that the parameters for the model featuring the DK category are also well recovered.

Real Data Analysis

Non-Cognitive Testing: Health Belief and the Common Sense Model of Diabetes Inventory

Using LLTM, we reanalyzed the data set that was used to illustrate IRT modeling for DK options in Ip et al. (2016). The purpose was to make a comparison with the previous analysis. Item responses collected from the Common Sense Model of Diabetes Inventory (CSMDI) and other measures were collected in a study of folklore beliefs about diabetes from $n=593$ participants with the disease (Grzywacz et al., 2011). An example item from the CSMDI is “Drinking lots of water helps to flush extra sugar out of the body.” Each of the 31 items in CSMDI contains three response categories: “Agree,” “Disagree,” and “Don’t Know (DK).”

The responses to the belief items were scored according to whether or not the response was consistent with the American Diabetes Association guidelines at the time (ADA, 2012). Because diabetes is a chronic disease that requires careful self-management, a patient’s belief system across different self-management domains has an important impact on how the disease progresses. Those beliefs that were consistent with the biomedical model, as defined by the ADA, were treated as “Correct.”

The cognitive operations included in the analysis were related to six different aspects of diabetes self-management: symptoms, information, causes, consequences, behavioral management, and medical management. An item can be “loaded” onto one or more of the cognitive operations. The Q-matrix specifying the item attributes is given in Appendix A.

For the DK option, we compared two models: Model I, in which a common shift term was used for all items that contained DK responses; and Model II, in which an item-specific shift term was used for each item that contained DK responses. Besides the item attributes, we included into the LLTM three personal covariates: gender, education level, and a measure of poverty. Education level contained three categories: “Did not finish high school” (reference category), “High school,” and “Above high school.” Poverty was a dichotomized variable: “Yes” and “No” (reference category).

In the CSMDI data set, there were only a small number of DK responses. Particularly, three items (17, 18, and 24; see Appendix A) had DK responses respectively at 9.3%, 2.4%, and 4.4%. A summary of the result of the analysis of the CSMDI data is shown in Table 4.

The estimates for cognitive operations (β parameters) only minimally change with the addition of person covariates. The ability estimate in the model represents the extent to which the person’s belief system is consistent with the biomedical model. The high positive value of symptoms coefficient seems to suggest that symptoms, as direct experiences that a patient must go through on a daily basis, are less likely to be aligned with the biomedical model. On the other hand, medical management coefficient is negative, suggesting that the patients in general knew about what they had been informed of, presumably from the healthcare system and other sources, regarding medicine and treatment of the disease.

Cognitive Test: KNOW Diabetes diabetes knowledge Test

Designed by a team that consisted of clinicians, endocrinologists, psychometricians, dietitians and nurses who specialize in diabetes care, the KNOW Diabetes diabetes knowledge test (DKT) contained 15 items for evaluating diabetes self-care knowledge in persons with diabetes. Response data had been collected from $n=227$ participants, with a known diagnosis of type 2 diabetes mellitus who were recruited from the Maryland/ Washington, D.C. and Chicago area during 2014-2016. All the items are presented in a multiple-choice format with four response options, one of which is DK. The DKT contains two branched-out questions about pill usage and insulin injection. Because only a fraction of the participants responded to the branched questions, in this analysis we excluded the two pill- and insulin-related questions so that responses from a total of 13 items were analyzed. Preliminary analysis using our proposed/current model/the model presented in this paper, showed that the DKT was both valid and reliable. A report documenting its psychometric properties is currently under preparation. To set up the LLTM for this example, we used only one attribute: namely, whether or not the item involves numeric knowledge. For example, the question “A good blood sugar in the morning before you eat anything is: A. 80-140, B. less than 150, C. less than 200, D. I don’t know” (correct answer is A) would be considered as having the attribute for numeric knowledge. The DKT and coding of the attribute are included in Appendix B.

The average percentage of DK among the 13 items was 10.3% (range 3.1-23.8%). The highest percentage of DK came from two questions (tie) respectively about the value of blood sugar after a meal and treatment of low blood sugar. On the other hand, on average, the correct percentage of responses across items was 75.4% (range 36-94%). The questions were self-administered by the participants, and perhaps because a DK option was provided, there were no recorded missing values in the responses.

In our analysis, three different procedures were used for scoring the DK response: DK as incorrect, as missing, and as a poset category as modeled by Equation (3)-(5). For parsimoniousness of the model for DK response, the reported poset model featured only a global shift parameter, δ , and not item-specific shifts. Table 5 compares the LLTM results of the three different schemes for coding DK.

In the model in which DK is coded as a poset category, the preprocessing indicated that DK should be coded as 1 across all items, implying that those who selected the DK option tended to have a higher level of ability compared to the collective group of participants who responded correctly or incorrectly. This is an interesting point that will be further considered in the Discussion section.

Table 5 shows the impact of the different approaches for handling DK. The first direct impact is on the intercept or offset. When DK is treated as incorrect, the intercept offset parameter is lower than in the other two coding schemes. Based on the offset value, it can be derived that on average, the probability of a correct response is $\exp(0.98)/[1+\exp(0.98)]=0.73$ when DK is treated as incorrect, and the probability is 0.82 for the other two schemes, given mean values of the covariates. The second impact is on the point estimates of the item attributes and person covariates in the LLTM. For example, the

numeric attribute is deemed statistically significant across the three coding schemes but their magnitudes are different. Individuals tend to have higher probabilities of correctly responding to items that have a numeric attribute. In DK as poset, everything else being equal, having a numeric attribute in the item leads to a probability of 0.85 of a correct response, compared to 0.82 of not having a numeric attribute. Age is significant in both DK as incorrect and DK as poset but not in DK as missing. For example, in DK as poset, younger age (< 45 years) was found to be highly significantly different from older age (> 65 years). The consistently positive values of the estimates across the three schemes imply that older age is associated with better performance on the test.

The third impact is on the standard error of the estimates. When DK is treated as incorrect, this assumption tends to enhance artificially the accuracy of the estimates so SE in Table for DK as incorrect is lower than for DK as missing. This means that it is more likely to see a higher number of significant results when DK is treated as incorrect. Of note, the standard errors using the poset model are also generally lower than the other two models, reflecting the fact that information has been gleaned from the DK response for informing the estimates. The reduction of standard errors is slightly more pronounced in the poset model than the other two models when covariates were not present (results not shown to save space). In other words, by treating DK as poset, one can still maintain better power to detect differences than by treating DK as missing.

Discussion

Analyzing Don't Know responses is a challenging task in educational and psychological testing. First and foremost, it is often not even clear if the DK response that a test taker provides represents a true or valid response. Several theories exist for understanding what a DK response actually means, and each theory has subsequent analytic implications. The most important distinction between these theories, may be considered as whether DK can be viewed as a valid response. Put differently, if DK is a valid response, then it is a true value; it is not a missing value or something that does not genuinely represent the respondent's cognitive status. If that is the case, then it would be important to treat DK as a legitimate response option and analyze it as such.

On the other hand, DK can also be an invalid value. Among theories that view DK as not genuinely reflecting the status of a respondent, the satisficing theory is most often cited. Krosnick and colleagues (1991, 2002) hypothesized that DK responses are the result of survey satisficing – that is, respondents settle for a satisfactory response that will appease the interviewer, rather than searching for an optimal response, which would require more cognitive work. Other possible theories include DK as a response style for some respondents (Bachman & O'Malley, 1984) and as a means for diplomatically and passively refusing to answer a question. The latter is especially likely for sensitive survey questions that involve racial ideology and controversial political subjects, for instance.

In this paper, we make the case that DK is a valid response. We argue that in many cases, for example, in low-stakes cognitive tests, DK is a legitimate response option and contains substantive information. This article demonstrates how DK responses for both cognitive and

non-cognitive tests can be modeled using a scientifically disciplined approach. Specifically, the LLTM forms the workhorse for inferring cognitive operations embedded in the DK responses. Under a unified LLTM modeling framework, the inclusion and analysis of the DK category allows information in DK to be gleaned and incorporated into the latent trait estimate. As demonstrated in the real data examples, the information gleaned from the DK response category is relatively weak compared to that from the directional responses, unsurprisingly. However, the accrual of such information, even if relatively weak, over multiple test items can offer a substantially more precise estimate for the construct of interest or for the cognitive operation underpinning the response process.

There are situations when DK is not considered a valid response or is viewed with suspicion about its possible use. In such cases, DK should be treated using other methods. There are several commonly used strategies which can be applied if this approach is taken. For example, the DK category can be coded as a directional response (e.g., collapsed into the Incorrect category), deleted in complete case analysis, or treated as a missing value. In the latter case, methods such as imputation (Rubin, 1987) can then be applied to the data set.

It needs to be emphasized that there are many real-world situations in which DK responses should be treated as valid and legitimate; methods such as imputation for missing values for DK would make little sense in those situations. For example, in a test such as the KNOW Diabetes DKT or a low-stakes survey such as career aspiration, within which there are likely to be scenarios that the respondent might not be entirely familiar with (Staff, Harris, Sabates, & Briddell, 2010), treating DK as missing or forcing it to take a directional value would lead to biased results. From a pragmatic perspective, the DK response avoids having the test taker guess at a response which may have its own inherent limitations. For example, in the DKT if the participant guesses correctly, then he or she may not receive the education that is required to address a knowledge deficit that should be important to the self-care process and would have been ‘flagged’ appropriately, that is as a knowledge deficit, by a DK response. Indeed, in many of these tests, objectivity is the reason a researcher wanted to integrate the DK option into the test in the first place. An added benefit of including DK into the response option in a test like KNOW Diabetes DKT, as evidenced in our result, is that missing responses could be minimized.

We have not attempted to model the underlying response psychology for DK response. As a reviewer pointed out, a DK response requires that the participant both not know and realize that they do not know, and that there could be alternative reasons for DK such as insufficient confidence, or to minimize effort to avoid committing oneself. In our analysis of the KNOW diabetes knowledge test data, the respondents that used DK were actually “rewarded” in the sense that their DK responses positively contributed to ability estimate. While using percentage correct as a metric, individuals that had at least one DK response and no DK were not very different (respectively 84.9% and 84.6%), treating DK as incorrect will likely downwardly bias the knowledge estimates for those who selected DK. For a low-stake test like KNOW, it is likely that those who used DK realized they don’t know and were willing to say they don’t know. It would be a “penalty” to this group if DK was treated as incorrect for their not guessing. The data appeared to suggest that a positive adjustment to their ability estimate (i.e., using the poset model for DK) was justifiable.

One way to address the response mechanism issue underlying DK is to collect more information such as measure of feeling of knowing (FOK; Koriat, 1993; Nelson, 1984; Reder, 1987). Especially with cognitive (knowledge) questions, it is possible that a participant feels that they know, but cannot retrieve the answer. In the case of KNOW Diabetes DKT, a participant may remember that he/she had heard about the proper range of blood sugar level in the morning from a physician but fails to recall the numbers. Measures of FOK can be used to distinguish between such memory lapse and a complete ignorance about the right sugar level. When new information about cognitive conditions are collected, it is likely that some corresponding response categories can be ordered alongside the other existing response categories while others may not be directly comparable to the existing categories. It can be argued that different response categories associated with different underlying mechanisms of responding to categories (e.g., DK, Know but cannot recall) will also likely constitute a poset. By properly structuring the response categories, the data can be handled by extension of the proposed poset-based method.

It is interesting to note that the subitem recoding scheme of poset also offers insight into the connection between treating DK as missing and the proposed LLTM. Operationally, the LLTM recodes the directional response as a subitem with DK as missing, and then the DK response as a separate subitem. Both subitems contribute to the estimation of ability. Without the DK subitem, the LLTM would be equivalent to simply treating DK as missing. In the proposed LLTM, the purpose is to extract additional information from the DK response for the same latent ability and cognitive operation as inferred by the directional response. Indeed, in the cognitive test example above, those who selected DK were found to contribute positively to the ability estimate compared to the collective group of participants that correctly or incorrectly responded to the question, although the contribution from the DK response was not as strong as the contribution from the Correct/Incorrect component of the model. It is also possible to model the DK (versus the directional responses) as measuring a latent trait distinct from the original latent ability of interest. As an example, in a social survey, the alternative latent trait could indicate an individual's propensity to stay neutral or have no opinion. One possibility is a joint model of the original construct together with neutrality propensity along the line of work proposed in Ip, Wang, De Boeck, & Meulders (2004). While the current LLTM does not directly investigate the alternative latent trait model, the poset approach offers a statistical framework for operationalizing other interesting LLTMs with DK option.

The proposed LLTM methodology can be further generalized in at least two directions. First, response categories, including DK, can be viewed as a partially ordered set (poset) – i.e., responses that do not completely follow a rank order, so the LLTM can be generalized to poset. This would represent a generalization of the work on poset in Zhang and Ip (2012) and Ip et al. (2016). A simple extension of the current application would be to ordinal response data (e.g., Agree, Neutral, and Disagree) with a separate DK option (Fig. 1(b)). In other words, the treatment of DK and other categories such as decline to respond can be generally applied to ordinal IRT models and related latent variable methods. A second direction is to generalize the method to non-directional categories in surveys and tests. For example, in political surveys, the non-directional neutral category is often analyzed as missing value even though it may represent a genuine value. Other non-directional

categories include omitted responses and categories such as “Refuse to Answer.” Finally, from a mathematical perspective, the extension of the DK analysis from IRT to LLTM illustrates the potential of the methodology to generalize to more complex latent variable models for DK responses. Currently, two pieces of research work are underway: analysis of DK responses in multidimensional IRT and analysis of poset, including responses with the DK option, in a longitudinal setting.

The current work has several limitations. First, the coding of DK as a directional response is different for different items, and without prior knowledge as to the direction the DK response contributes to, we need to preprocess the data for the determination of direction. The proposed method involves a two-stage procedure that may not be optimal because the ability estimate of theta in the first stage may be dependent on the estimate in the second stage. The procedure can be tedious as well when the number of items is large. Second, the LLTM is based on the Rasch model, which some may view as too restrictive. For example, some might prefer adding an item-level discrimination parameter to the model. Although the recoding scheme still works, PROC GLIMMIX can no longer be used; it is possible to use PROC NLMIXED for estimation. In either case, it is important that the Q-matrix is well specified and properly captures the cognitive operations embedded in the response to the stimuli. Misspecification of the Q-matrix can lead to large biases (Baker, 1993; MacDonald, 2014). Evidence from simulation studies showed that with a correctly specified Q-matrix, these programs can both accurately and efficiently recover the LLTM (MacDonald & Kromrey, 2011). Another limitation of the method is that DK is assumed to be a valid response. When this does not hold, then the information the DK category provides may be misleading, and the result would be biased. A researcher therefore must use other information such as the content of the test and the context under which the test is conducted to make a judgment about the legitimacy of the DK response.

Conclusions

It has been suggested that The Don't Know (DK) response contains important information that is often overlooked in interpreting the data collected from cognitive and non-cognitive tests.

Direct psychometric modeling efforts for DK responses have been limited to-date. The LLTM presented in this paper, with some practical examples relevant to the diabetes field, offers a potentially practical and easy-to-implement tool for making inferences about cognitive operations when a test, cognitive or otherwise, contains the DK option. We have shown how the proposed LLTM offers a potentially practical and easy-to-implement tool for making inferences about cognitive operations when a test, cognitive or otherwise, contains the DK option. Our results show that weak information regarding the targeted latent trait can be inferred from DK responses. Further research is needed to validate this application of LLTM.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment:

This research is supported by National Science Foundation grant SES-1229549 and SES-1424875 awarded to Edward H. Ip and National Institutes of Health grants UL1TR001420 through the Wake Forest Clinical and Translational Science Institute.

Appendix A: Common Sense Model of Diabetes Inventory (CSMDI). Reverse coded items are coded as correct when a negative response (Disagree) is registered.

Item no.	description	S*	I*	Ca*	Co*	M*	MM*
1	Feeling nervous is a sign of low blood sugar	1	0	0	0	0	0
2	Blood sugar will go up if you eat too many white foods (reverse)	1	0	0	0	0	0
3	People with diabetes have tingling in their feet due to high blood sugar	1	0	0	1	0	0
4	Falling down is a sign of diabetes (reverse)	1	0	0	1	0	0
5	Having to go to the bathroom often at night is caused by diabetes	1	0	0	1	0	0
6	Diabetes makes people feel thirsty all the time	1	0	0	1	0	0
7	Family members with diabetes are good sources of diabetes information (reverse)	0	1	0	0	0	0
8	People with diabetes understand their disease better than their doctors (reverse)	0	1	0	0	0	0
9	People could better control their diabetes if they were given the right information	0	1	0	0	0	0
10	Weight does not cause diabetes because thin people also get diabetes (reverse)	0	0	1	0	0	0
11	Diabetes can't be hereditary because not everyone in a family gets it (reverse)	0	1	1	0	0	0
12	Some people get diabetes because they ate too many sweets when they were young (reverse)	0	1	1	0	0	0
13	Everyone is born with diabetes but it develops at different times for different people (reverse)	0	1	1	0	0	0
14	Being overweight makes people get diabetes	0	1	1	0	0	0
15	Diabetes runs in families	0	1	1	0	0	0
16	Diabetes causes high blood pressure (reverse)	0	0	0	1	0	0
17**	It is difficult for people with diabetes when they have a full-time job	0	0	0	1	0	0
18**	Diabetes has serious financial consequences	0	0	0	1	0	0
19	Diabetes makes it difficult for your body to fight infection	0	0	0	1	0	0
20	Drinking lots of water helps to flush extra sugar out of the body (reverse)	0	0	0	1	1	0
21	Stress makes your blood sugar go up	1	0	0	1	1	0
22	Managing the size of each meal helps control diabetes	0	0	0	0	1	0
23	The only thing people with diabetes need to know is to stay away from sweets (reverse)	0	0	0	1	1	0
24**	Doing household chores is enough exercise for someone who has diabetes	0	0	0	0	1	0

Item no.	description	S*	I*	Ca*	Co*	M*	MM*
25	The body processes sugar in fruits and vegetables differently than sugar in sweets and starches	0	0	0	0	1	0
26	Blood sugar often goes up and down for no reason (reverse)	1	0	0	0	1	0
27	Taking extra medication helps to manage high blood sugar (reverse)	0	1	0	0	0	1
28	People should adjust their diabetes medication depending on how they feel (reverse)	0	1	0	0	0	1
29	Low blood sugar can be managed by adjusting medication (reverse)	0	0	0	0	0	1
30	Medical treatment cures diabetes (reverse)	0	0	0	0	0	1
31	Taking extra medication makes it okay to eat something sweet (reverse)	0	0	0	1	0	1

* S=Symptoms, I=Information, Ca=Cause, Co=Consequences, M=Management, MM=Medical Management.

** Items #17,18, and 24 contain DK responses, and the shift parameters are respectively represented by δ_1, δ_2 , and δ_3 .

Appendix B: Diabetes Knowledge Test

Please answer the following questions, checking the best answer.

If you do not know the answer, check “Do not know.”

1. A person with diabetes should test his or her blood sugar at home at least:
 - a. One to two times per day*
 - b. Once a week
 - c. Once a month
 - d. Do not know
2. Checking blood sugar at home:
 - a. Provides me with information on my diabetes control
 - b. Helps me understand the effect of physical activity and food on my blood sugar
 - c. Both a and b*
 - d. Do not know
- #3. A good blood sugar in the morning before you eat anything is:
 - a. 80-140*
 - b. less than 150
 - c. less than 200
 - d. Do not know
- #4. The highest a blood glucose should be two hours after a meal or at any other time of the day is:

- a. 120
- b. 180*
- c. 250
- d. Do not know
- #5. The A1C blood test gives you your average blood glucose in the past 3 months. The A1C target or goal for most people with diabetes is less than:
- a. 7*
- b. 8
- c. 9
- d. Do not know
6. High blood sugar may make you:
- a. Feel shaky and nervous
- b. Feel full of energy
- c. Have increased urination, thirst and hunger*
- d. Do not know
- #7. Your blood sugar is greater than 300 for two or more tests in one day. You should:
- a. Call your doctor*
- b. Go to bed
- c. Stop taking diabetes medication
- d. Do not know
8. It is Saturday morning and your meter reads “High” and you feel sick. You should:
- a. Call your doctor on Monday to schedule an appointment
- b. Wait until your next scheduled appointment with your doctor in one month
- c. Go to the hospital emergency room*
- d. Do not know
9. Low blood sugar may make you feel:
- a. Shaky and nervous*
- b. Full of energy
- c. Itchy
- d. Do not know

10. Low blood sugar should **not** be treated with:
- ½ cup of diet soda*
 - 1 cup skim milk
 - 1 tablespoon of jelly, syrup, sugar or honey
 - Do not know
11. Fruit juice, soda, and sweet tea make blood sugar:
- Go down
 - Go up*
 - Stay the same
 - Do not know
12. Foods high in carbohydrates such as bread, potatoes, fruit, and milk make blood sugar:
- Go down
 - Go up*
 - Stay the same
 - Do not know
13. It is recommended that a person with diabetes:
- Eats 3 meals per day and eats less food*
 - Drinks plenty of natural fruit juice and fruit shakes
 - Avoids all foods that contain carbohydrates
 - Do not know

* Key

Coded as having numeric attribute

References

- ADA United States. (2012). Standards of medical care in diabetes. *Diabetes Care*, 35 (Suppl. 1), S11–S63. Retrieved from http://care.diabetesjournals.org/content/diacare/suppl/2016/12/15/40.Supplement_1.DC1/DC_40_S1_final.pdf DOI: 10.2337/dc12-s011 [PubMed: 22187469]
- Bachman JD, & O'Malley PM (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, 48, 491–509. 10.1086/268845
- Baker F (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement*, 17, 201–210. 10.1177/014662169301700301
- Bishop GF (2005). *The illusion of public opinion*. Lanhan, MD: Rowman & Littlefield Publisher, Inc.
- Chen Y-H, MacDonald G, & Leu Y-C (2011). Validating cognitive source of mathematical item difficulty: Application of LLTM to fraction conceptual items. *The International Journal of Educational and Psychological Assessment*, 7, 74–93.

- De Boeck P, & Wilson M (2004). Explanatory item response models: A generalized linear and nonlinear approach. New York: Springer.
- Dimitrov DM & Raykov T (2003). Validation of cognitive structures: A structural equation modeling approach. *Multivariate Behavioral Research*, 38, 1–23. DOI: 10.1207/S15327906MBR3801_1 [PubMed: 26771122]
- Draney K, & Wilson M (2008). A LLTM approach to the examination of teachers' ratings of classroom assessment tasks. *Psychological Science Quarterly*, 50, 417–432.
- Embretson S, & Wetzel D (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, 11, 175–193. 10.1177/014662168701100207
- Embretson SE (1995). A measurement model for linking individual learning to process and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement*, 32, 277–294. DOI: 10.1111/j.1745-3984.1995.tb00467.x DOI:
- Feick LF (1989). Latent class analysis of survey questions that include don't know responses, *Public Opinion Quarterly*, 53, 525–47. 10.1086/269170
- Fischer GH (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359–374. 10.1016/0001-6918(73)90003-6
- Fischer GH (1995). The linear logistic test model In Fischer GH & Molenaar IW (Eds.), *Rasch models: Foundation, recent developments, and applications* (pp. 131–156). New York: Springer-Verlag.
- Gorin J (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, 43, 351–373. DOI: 10.1111/j.1745-3984.2005.00020.x DOI:
- Grzywacz J, Arcury TA, Ip EH, Chapman C, Kirk JC, Bell RA, ... Quandt SA (2011). Older adults' common-sense models of diabetes: A multi-method approach illustrating the importance of viewing health beliefs holistically. *American Journal of Health Behavior*, 35, 318–353. DOI: 10.5993/AJHB.35.3.6 [PubMed: 21683021] DOI:
- Hohensinn C, Kubinger KD, Reif M, Holocher-Ertl S, Hkorramdel L, & Frebort M (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychological Science Quarterly*, 50, 391–402.
- Ip EH, Wang Y, De Boeck P, & Meulders M (2004). Locally dependent latent trait model for polytomous responses with application to inventory of hostility. *Psychometrika*, 69, 191–216. DOI: 10.1007/BF02295940 DOI:
- Ip EH, Smits D, & De Boeck P (2009). Locally dependent linear logistic test model with person covariates. *Applied Psychological Measurement*, 33, 555–569. 10.1177/0146621608326424
- Ip EH, Chen S, Quandt S (2016). Analysis of multiple partially ordered responses to belief items with don't know option. *Psychometrika*. 81, 483–505. DOI: 10.1007/s11336-014-9432-y [PubMed: 25479822]
- Koriat A (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100, 609–639. 10.1037/0033-295X.100.4.609 [PubMed: 8255951]
- Krosnick JA (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213–236. DOI: 10.1002/acp.2350050305 DOI:
- Krosnick JA, Holbrook AL, Berent MK, Carson RT, Hanemann WM, Kopp RJ, Conaway M (2002). The impact of 'no opinion' response options on data quality: Non-attitude reduction or an invitation to satisfy? *Public Opinion Quarterly*, 66, 371–403. 10.1086/341394
- Kubinger KD (2008). On the revival of the Rasch-model based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychological Science Quarterly*, 50, 311–327.
- Kubinger KD (2009). Application of the linear logistic test model in psychometric research. *Educational and Psychological Measurement*, 69, 232–244. 10.1177/0013164408322021
- Lord FM (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.
- MacDonald G, & Kromrey J (2011). Linear logistic test model: Using SAS® to simulate the decomposition of item difficulty by algorithm, sample size, cognitive component and time to

- convergence. Proceedings of the American Statistical Association's Joint Statistical Meeting, Social Statistics Section [CD-ROM], Miami, FL.
- MacDonald GC (2014). The performance of the linear logistic test model when the Q-matrix is misspecified: A simulation study. Graduate Theses and Dissertations. Retrieved from <http://scholarcommons.usf.edu/etd/5065>
- Mair P, Hatzinger R, & Mair MJ (2014). eRm: Extended Rasch modeling [Computer software]. R package version 0.15–4. Retrieved from <http://CRAN.R-project.org/package=eRm>
- Mayer R, Larkin J, & Kadane P (1984). A cognitive analysis of mathematical problem solving In Sternberg R (Ed.) Advances in the psychology of human intelligence (Vol. 2, pp. 231–273). Hillsdale, NJ: Erlbaum.
- Nelson TO (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. Psychological Bulletin, 95, 109–133. 10.1037/0033-2909.95.1.109 [PubMed: 6544431]
- Rasch G (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Danish Institute for Educational Research.
- Reder LM (1987). Strategy selection in question answering. Cognitive Psychology, 19, 90–138. 10.1016/0010-0285(87)90005-3
- Reise SP & Waller NG (2009). Item response theory and clinical measurement. Annual Review of Clinical Psychology, 5, 27–48. DOI: 10.1146/annurev.clinpsy.032408.153553
- Rijmen F, & De Boeck P (2002). The random weights linear logistic test model. Applied Psychological Measurement, 26, 271–285. 10.1177/0146621602026003003
- Rijmen F, Tuerlinckx F, De Boeck P, & Kuppens P (2003). A nonlinear mixed model framework for item response theory. Psychological Methods, 8, 185–205. 10.1037/1082-989X.8.2.185 [PubMed: 12924814]
- Rubin DB (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.
- Staff J, Harris A, Sabates R, & Briddell L (2010). Uncertainty in early occupational aspirations: Role exploration or aimlessness. Social Forces, 89, 659–684. doi: 10.1353/sof.2010.0088
- Tatsuoka KK (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 345–354. DOI: 10.1111/j.1745-3984.1983.tb00212.x
- Zeuch N, Holling H, & Kuhn JT (2011). Analysis of the latin square task with linear logistic test models. Learning and Individual Differences, 21, 629–632. 10.1016/j.lindif.2011.03.004
- Zhang Q, Ip EH (2012). Generalized linear model for partially ordered data. Statistics in Medicine, 31, 56–68. doi: 10.1002/sim.4318 [PubMed: 22086670]

(a)

(b)

Fig. 1.

(a) Directional and non-directional responses respectively for dichotomous response categories and DK, (b) Directional and non-directional responses respectively for polytomous response categories, and DK and Refuse to respond.

Table 1.

Poset Conditional Coding Scheme for an Item with Response Options Correct, Incorrect, and Don't Know (DK) (Fig. 1)

Response Category in Poset Structure Y	Subitem 1 $Y^{(1)}$	Subitem 2 $Y^{(2)}$
<i>Correct</i>	1	0
<i>Incorrect</i>	0	0
<i>DK</i>	NA	1**

* Subitem 1 and Subitem 2 respectively correspond to Model A (Equation (3)) and Model B (Equation (4))

** Assume DK positively contributes to ability compared to the directional categories

Table 2.

Results from Simulation Study 1 (SS1)

Bias	Set A (18 items)		Set B (18 items)		Set A+B (36 items)	
	N=200	N=1000	N=200	N=1000	N=200	N=1000
β_0	-0.003	0.006	0.014	0.006	-0.004	0.003
β_1	-0.015	-0.004	-0.017	-0.005	-0.008	-0.002
β_2	0.007	0.001	-0.015	-0.004	0.002	0.002
β_3	-0.005	-0.006	-0.010	-0.006	-0.002	-0.005
β_4	-0.012	-0.005	-0.024	-0.005	-0.017	-0.005
β_5	-0.008	-0.007	0.005	0.007	-0.001	0.000
RMSE						
β_0	0.013	0.003	0.022	0.005	0.009	0.002
β_1	0.010	0.002	0.011	0.002	0.004	0.001
β_2	0.009	0.002	0.010	0.002	0.005	0.001
β_3	0.012	0.002	0.009	0.002	0.004	0.001
β_4	0.008	0.001	0.007	0.001	0.003	0.001
β_5	0.006	0.001	0.007	0.002	0.003	0.001

Table 3.

Results from Simulation Study 2 (SS2)

Bias	Set 1 (18+3 items)		Set 2 (18+3 items)		Set 3 (36+3 items)		Set 4 (36+3 items)		Set 5 (36+6 items)	
	N=200	N=1000	N=200	N=1000	N=200	N=1000	N=200	N=1000	N=200	N=1000
β_0	-0.003	0.005	0.016	0.006	-0.003	0.002	-0.003	0.004	-0.002	0.002
β_1	-0.015	-0.003	-0.018	-0.004	-0.008	-0.001	-0.009	-0.002	-0.009	-0.001
β_2	0.005	0.002	-0.015	-0.004	0.002	0.002	0.002	0.002	0.002	0.003
β_3	-0.004	-0.004	-0.011	-0.006	-0.002	-0.004	-0.003	-0.005	-0.004	-0.004
β_4	-0.014	-0.005	-0.023	-0.004	-0.018	-0.005	-0.016	-0.004	-0.017	-0.004
β_5	-0.007	-0.006	0.001	0.005	0.000	0.001	-0.003	0.000	-0.002	0.000
δ_1	0.002	-0.001	-	-	0.007	0.002	-	-	0.005	0.002
δ_2	-0.004	-0.012	-	-	0.007	-0.008	-	-	0.005	-0.010
δ_3	-0.005	0.018	-	-	-0.001	0.014	-	-	-0.001	0.015
δ_4	-	-	-0.028	-0.007	-	-	-0.019	0.000	-0.019	0.000
δ_5	-	-	-0.005	0.005	-	-	-0.006	0.006	-0.005	0.006
δ_6	-	-	-0.043	-0.018	-	-	-0.047	-0.021	-0.047	-0.020
RMSE										
β_0	0.014	0.003	0.022	0.005	0.009	0.002	0.009	0.002	0.009	0.002
β_1	0.010	0.002	0.012	0.002	0.004	0.001	0.004	0.001	0.004	0.001
β_2	0.009	0.001	0.010	0.002	0.005	0.001	0.005	0.001	0.005	0.001
β_3	0.012	0.002	0.010	0.002	0.004	0.001	0.004	0.001	0.004	0.001
β_4	0.008	0.001	0.007	0.001	0.003	0.001	0.003	0.001	0.003	0.001
β_5	0.007	0.001	0.007	0.002	0.003	0.001	0.003	0.001	0.003	0.001
δ_1	0.039	0.006	-	-	0.033	0.005	-	-	0.033	0.005
δ_2	0.042	0.008	-	-	0.037	0.007	-	-	0.038	0.007
δ_3	0.050	0.008	-	-	0.047	0.008	-	-	0.047	0.008
δ_4	-	-	0.052	0.010	-	-	0.047	0.008	0.047	0.009
δ_5	-	-	0.057	0.008	-	-	0.047	0.008	0.047	0.008
δ_6	-	-	0.061	0.009	-	-	0.045	0.007	0.045	0.007

Table 4.

Summary of Results of Analysis of the Common Sense Model of Diabetes Inventory (CSMDI) for Linear Logistic Test Model (LLTM) (N=593)

Effect	LLTM		LLTM with Person Covariates	
	Model I*	Model II*	Model I*	Model II*
Intercept	1.42 (0.06)	1.42 (0.06)	1.93 (0.09)	1.93 (0.09)
Symptom	0.42 (0.05)	0.42 (0.05)	0.41 (0.05)	0.41 (0.05)
Information	-0.27 (0.05)	-0.27 (0.05)	-0.26 (0.05)	-0.26 (0.05)
Cause	-0.94 (0.06)	-0.94 (0.06)	-0.94 (0.06)	-0.94 (0.06)
Consequence	-0.50 (0.05)	-0.50 (0.05)	-0.49 (0.05)	-0.50 (0.05)
Management	-0.54 (0.05)	-0.54 (0.05)	-0.55 (0.05)	-0.54 (0.05)
Medical Management	-1.51 (0.05)	-1.51 (0.05)	-1.52 (0.05)	-1.52 (0.05)
δ (model I)	-3.93 (0.11)	-3.33 (0.15)	-3.97 (0.11)	-3.38 (0.15)
δ_1 (model II)				
δ_2 (model II)		-4.80 (0.28)		-4.86 (0.29)
δ_3 (model II)		-4.12 (0.21)		-4.13 (0.21)
Gender			-0.13 (0.06)	-0.13 (0.06)
Education (>HS)			-0.59 (0.07)	-0.59 (0.07)
Education (HS)			-0.40 (0.07)	-0.40 (0.07)
Poverty			-0.18 (0.06)	-0.18 (0.06)

Model I: a common shift term δ used for the DK model

Model II: an item-specific shift term δ_i , $i=1,2,3$, used for the DK model

Table 5.

Summary of Analysis Results for Diabetes Knowledge Test (N=227)

Effect	Estimate	SE	p-value
DK as incorrect			
Intercept	0.98	0.10	<0.001
Has numeric attribute	0.20	0.09	0.03
Age (<45) *	0.73	0.20	<0.01
Age (45-65) *	0.49	0.13	<0.01
Education *	-0.62	0.27	0.02
Length of diabetes (<1y) *	-0.23	0.19	0.23
Length of diabetes (1-6y) *	-0.25	0.15	0.09
DK as missing			
Intercept	1.51	0.11	<0.001
Has numeric attribute	0.30	0.11	0.008
Age (<45) *	0.31	0.21	0.13
Age (45-65) *	0.13	0.13	0.31
Education *	-0.14	0.30	0.64
Length of diabetes (<1y) *	0.13	0.21	0.55
Length of diabetes (1-6y) *	0.03	0.15	0.83
DK as poset			
Intercept	1.51	0.10	<0.001
Has numeric attribute	0.18	0.09	0.03
Age (<45) *	1.23	0.17	<0.001
Age (45-65) *	0.45	0.19	0.01
Education *	-0.40	0.26	0.12
Length of diabetes (<1y) *	-0.04	0.19	0.81
Length of diabetes (1-6y) *	-0.11	0.14	0.42

* Reference categories are age > 65, education not less than high school, length of diabetes > 6 years.