

Running head: ITEM RESPONSE TIME AND DISTRACTOR ANALYSIS

Including Item Response Time in a Distractor Analysis
via Multivariate Kernel Smoothing

J. Patrick Meyer and Steven L. Wise

James Madison University

March 24, 2005

Paper presented at the 2006 meeting of the National Council on Measurement in Education, San Francisco, CA.

Abstract

Response time models and distractor analysis are two methods for obtaining information about construct representation and developing a better understanding of examinee cognition. This study describes a kernel smoothing approach for including item response time in a distractor analysis. The method was applied to a 60 item web-based information literacy test. Option characteristic surfaces for four exemplar items were used to demonstrate the benefit of including response time in a distractor analysis.

Including Item Response Time in a Distractor Analysis via Multivariate Kernel Smoothing

Time is an important component in understanding the way an examinee processes an item and selects a response. Response time analysis has revealed that the correct and incorrect response may not be processed in the same manner (Thissen, 1983) and that examinees may engage in different types of response behavior (Schnipke & Scrams, 1997) or exhibit different levels of motivation (Wise & Kong, 2005). Indeed, Embretson (1999) noted the importance of response time in supporting construct representation, which concerns the meaning of test scores. In her cognitive design system framework (Embretson, 1999), measurement models that account for response speed and accuracy may be used to gather empirical evidence about item characteristics and their relationship to the cognitive theory underlying a test. Several models of response time and accuracy have been developed (Roskam, 1997; Thissen, 1983; Van Breukelen 2005; Verhelst, Verstralen, & Jansen, 1997), and these models make different assumptions about the speed-accuracy trade off function (SAF). Although such assumptions facilitate the explanatory power of a model, they limit its exploratory power. An exploratory analysis may be more useful when little is known about the relationship between time, item features, and examinee cognition. Multivariate kernel smoothing, described below, is an approach that is particularly well-suited for exploratory analysis.

Another way to obtain evidence in support of construct representation is through a distractor analysis. Researchers have demonstrated that incorrect response options, or distractors, may be used to improve inferences about examinee cognition. Thissen (1976, 1993) demonstrated that the nominal response model (NRM; Bock, 1972) can improve ability estimation and inform researchers about the cognitive processes that underlie item responses. Levine and Drasgow (1983) showed that the choice of incorrect response options was often

related to ability. At low levels of ability, for example, one wrong option could be attractive, but at high levels, another type of incorrect option could be more attractive. Their findings suggest that selecting the wrong response is the result of a non-random process. Barton and Huynh (2003) used multiple-choice distractors to compare patterns of errors among examinees that needed or did not need an oral reading accommodation. They found statistically significant, albeit unsubstantial, support for different cognitive processes among these two groups. Each of these studies demonstrated the value of using all response options to improve inferences about examinee cognition.

The purpose of this study is to describe a method for adding response time to a distractor analysis and to demonstrate the benefit of doing so using examples from a web-based information literacy test.

Kernel Estimation of Option Characteristic Curves

A number of parametric models for estimating trace lines or option characteristic curves (OCCs) are possible when items are scored in one of several categories (see Drasgow, Levine, Tsien, Williams, & Mead, 1995). Of these models, the most unconstrained are the nominal response model (Bock, 1972) and the more general multiple choice model (Thissen & Steinberg, 1997), which do not require the categories to be ordered in any manner. Ramsay (1991, 1997) described an even more flexible approach for item characteristic curve estimation. He used a nonparametric kernel estimator to estimate option characteristic curves. In particular, he used the Nadaraya-Watson estimator,

$$P_j(x = 1 | \theta) = \frac{\sum_{i=1}^N K\left(\frac{\theta - \theta_i}{h_1}\right) x_{ij}}{\sum_{i=1}^N K\left(\frac{\theta - \theta_i}{h_1}\right)}, \quad (1)$$

where $K(\cdot)$ is the kernel function, h_1 is the smoothing parameter, θ is an evaluation point along the ability dimension, and θ_i is an examinee's normalized test score. To obtain OCC estimates, the x 's are dummy coded to indicate whether or not a particular response option was endorsed and Equation 1 is computed for each option. An advantage of using this type of estimator is that additional dimensions may be easily incorporated.

Kernel Estimation of Option Characteristic Surfaces

Kernel smoothing provides a flexible approach for exploring the relationship between several dimensions and the probability of a correct response. Commonly in educational measurement, the dimensions are measured via item responses, and in the case of simple structure, each dimension is measured by a subset of items such that each item may contribute to only one subtest score. As mentioned in Ramsay (1991), each subtest score may be used in a multivariate kernel estimator to obtain item characteristic surfaces. In the case of two dimensions, the estimator is given by,

$$P_j(\theta, \lambda) = \frac{\sum_{i=1}^N K\left(\frac{\theta - \theta_i}{h_1}, \frac{\lambda - \lambda_i}{h_2}\right) x_{ij}}{\sum_{i=1}^N K\left(\frac{\theta - \theta_i}{h_1}, \frac{\lambda - \lambda_i}{h_2}\right)}, \quad (2)$$

where the univariate kernel function in Equation 1 is replaced by a product

kernel, $K(u, v) = K(u)K(v)$; θ and λ are points along the two dimensions; θ_i and λ_i are normal score transformations of an examinee's subtest scores; and h_1 and h_2 are the smoothing parameters for each dimensions, respectively. The smoothing parameters h_1 and h_2 are specified independently and may or may not be equal. As in the case of the univariate estimator, Equation 2 may be obtained for each response option.

To incorporate item response time into estimation of the item characteristic surface, Meyer (2006) used a normal score transformation of the raw test score for one dimension and a robust standardization of item response time for the other dimension. Specifically, examinee ability, θ_i , is estimated by removing the studied item from the raw test score, $R_i = \sum_{j=1}^p x_{ij} - x_{ij}$.

The rest scores, R_i , are then ranked with ties broken randomly and converted to normal scores, $\theta_i = \Phi^{-1}\left(\frac{r_i}{N+1}\right)$, where r_i is the rank of R_i . These transformed rest scores form the basis of the ability dimension.

To include item response time as the second dimension, let t_{ij} be the observed response time of person i to item j , and let the item time deviation score be $\lambda_{ij} = [t_{ij} - \gamma(t_j)] / IQR(t_j)$, where $\gamma(t_j)$ and $IQR(t_j)$ are the median and interquartile range of the item response time across examinees, respectively. Although Meyer (2006) used this robust measure of deviation it is not a requirement for the estimator. Alternatively, response time itself may be used as well as other transformations such as the natural log of response time.

Method

Instrumentation

The Information Literacy Test (ILT) is a 60-item multiple-choice web-based based test. It was designed to measure Standards 1, 2, 3, and 5 of the *ACRL Information Literacy Competency Standards for Higher Education Standards* (ACRL, 2000). Reliability for this administration of the ILT was acceptable (Cronbach's alpha = 0.88). Dimensionality of the item scores was checked using DIMTEST (Stout, 1987). Stout's T indicated that the ILT was unidimensional (T = 1.1726, p = 0.12).

Participants

A sample of students ($N = 524$) at a southeastern community college responded to the ILT. The sample consisted of female (56%) and male college students. Caucasians comprised the largest race/ethnicity group (85%), while African Americans (4%), Asians (3%), Hispanics (2%), and Native Americans (1%) were represented in smaller numbers.

Procedure

Data from the spring 2004 administration of the ILT were analyzed. An item and response time analysis was first conducted. Next, option characteristic curves (OCCs) were estimated using Equation 1. Graphs of these curves were then compared to option characteristic surfaces (OCSs) estimated using Equation 2.

Although several choices of kernel are possible (see Härdle, Müller, Sperlich, & Werwatz, 2004, p. 41) the Gaussian kernel,

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right),$$

was used for both dimensions, θ and λ . The kernel was evaluated at 31 evenly spaced points between -2.5 and 2.5. Normal rest scores were computed by ranking the rest scores and converting them to normal scores. Item response times were converted to response time deviation scores (although these scores may still be loosely referred to as response time hereafter). Smoothing parameters were estimated using Scott's rule $h_k = \hat{\sigma}_k n^{-1/(d+4)}$ (Scott, 1992, p. 152), where $\hat{\sigma}_k$ is the standard deviation of the score for dimension k , and d is the number of dimensions. Estimation was implemented with functions written in R Version 2.0.1 (R Development Core Team, 2004).

Results

A classical item analysis indicated that item difficulty, \hat{p} , ranged from .05 to .94 with a mean of 0.69 and a standard deviation of 0.22. The point biserial correlation, r_{pbis} , ranged from -.04 to .60 with a mean of 0.32 and a standard deviation of 0.12.

Analysis of item response times indicated that the median item response time, $\gamma(t_j)$, ranged from 7 to 52 seconds with a mean of 20.92 and a standard deviation of 11.46. The point biserial correlation, $\hat{\rho}(item, \lambda_i)$, computed between the item score and item time deviation score ranged from -.21 to .34 with a mean of .08 and a standard deviation of .11. Principal components analysis of the item response times revealed a single dominant component that accounted for 30% of the variance.

Four items were selected from the ILT to demonstrate the benefit of including item response time in a distractor analysis. These items were selected because they represented four different types of surfaces that emerged from the analysis of the entire test. Descriptive statistics for these items are shown in Table 1.

OCC and OCS Features

Figure 1 shows the OCCs for item A, which is an easy item that discriminates well between high and low ability examinees. Response option three appears to be the most attractive incorrect answer among examinees with an ability level below about -2.0. The fourth response option is the second most probable incorrect answer up to an ability level of about -2.0 at which point option 1 (which is correct) becomes more likely. At ability levels above 0, only the correct answer is endorsed.

The OCSs for item A, which incorporate response time information, are shown in Figure 2. As shown for option two, item A also discriminates well between fast and slow responding

examinees. However, this is only true for examinees with ability levels below about 0. The interesting feature of this item was that low ability examinees who respond slowly (and presumably think about the item more) tend to get it correct. Conversely, low ability examinees who answer the item hastily, tend to get it incorrect. The difference between these two groups of low ability examinees suggests that the item was processed differently or that there was a different level of engagement with the item. Item A demonstrates that item response time does not always provide a great deal of information about the selection of incorrect responses. Option three was the most probable incorrect option for low ability examinees responding faster than the median item response time.

Item B is also an easy item but it does not discriminate very well. The OCCs for this item show that the correct answer is typically the most endorsed item (see Figure 3). Among the incorrect responses, however, option one is most likely at low levels of ability. For this particular item, the OCCs do not really provide much additional information about the item. When item response time was incorporated, however, quite a different picture emerged.

The OCS for the correct option (option 2) shows that high ability examinees tend to get the item correct, regardless of the amount of time spent on the item (see Figure 4). Comparing the surfaces for options, one, two, and three suggests three different groups of low ability examinees: Those who respond quickly tended to answer option one; those who responded at about the median time tended to select option two, the correct answer; and those who spent a lot of time on the item were likely to select option three. This pattern suggests substantial differences in the processes that low ability examinees use to respond to the item. Indeed, the item content further supports this notion. Item B addressed knowledge about identifying the correct source for locating the definition of Huntington's chorea. If an examinee thought that this

term was related to human anatomy, the first option was the most plausible distractor and perhaps the one that required the least amount of thought. If an examinee thought this term was a person's name, then option three was the most plausible distractor. Low ability examinee's may have spent time debating the plausibility of this distractor versus the correct response option, yet only the correct response was plausible if this term was recognized as a medical disorder.

Item C was somewhat harder than the previous two items with a fair amount of discrimination (see Figure 5). Option three was consistently the most likely incorrect response, although among a very small range of low ability examinees, option two was most probable. Incorporating item response time did not yield much more information for the correct response (see Figure 6). However, there were clear differences between the selected distractor and item response time. Low ability examinees who spent a large amount of time on the item tended to select option two, but those who responded quickly tended to select option three. This finding was particularly true at extremely low levels of ability. Again, these OCSs suggest that low ability examinees process or engage the item differently. In terms of item content, this item required examinees to read a mock web page and extract the information needed to answer the item. Given the amount of reading this item required, fast responding low ability examinees seemed to only be guessing. In contrast, low ability examinees who responded slowly were likely to select option two. Although this option is incorrect, it shared some features with the correct response. Namely, option two and the correct option, and only these two options, concerned the relationship between garlic and cancer inhibition, which was the gist of the item.

Finally, Item D was the hardest and most discriminating item of those discussed herein. Figure 7 shows that option two was the most likely response among low ability examinees up to an ability level of about -1.5. From this point until about -0.9, option one was most likely. The

correct answer was the most probable response at higher ability levels. The OCCs demonstrated that ability alone does a good job in distinguishing between the selected distractors, unlike the previously discussed items. This pattern was also evident in the OCSs (see Figure 8). Adding item response time to the analysis shed additional light on examinees. For example, option two was the most likely response for fast responding examinees at all ability levels as well as for fast to moderately slow responding low ability examinees. The display of the item content suggests an explanation for this pattern. Item D related to search records obtained from an electronic database such as ERIC. Each response option was a different search record. Only the first two records (the first two response options) appear on the scrollable screen when the item is first presented. Of these two options, option two is more correct. A fast responding examinee may not have recognized the need to scroll the web browser to see the other two records. The moderately slow responding low ability examinees may not have scrolled the display but may have required more time to sort through the information in these two distractors. Examinees that saw all four records would necessarily have spent more time on the item and actually have seen the correct option, although simply seeing the option was not sufficient for answering the item correctly. There were other items on the ILT that required the display to be scrolled to see the entire item content. However, item D was the only one in which scrolling was required to see all of the response options.

Discussion

Example items from the ILT were used to demonstrate that including item response time in a distractor analysis provides additional insight into examinee cognition and the process of answering an item. Distractors that appear to function similarly or without distinction across ability may in fact perform quite differently from each other and reveal different groups of examinees when time is incorporated into the analysis. This information may be used as

evidence of construct representation, or it may be used to develop distractors that better discriminate between the various cognitive errors an examinee may make when responding to an item. The exploratory nature of this multivariate kernel smoothing procedure (Equation 2) is helpful when a researcher has no preexisting notion of the underlying cognitive process or trait manifest in item response times. By comparing OCSs and examining item content, meaningful patterns may emerge that inform researchers about examinee cognition. A stronger and more confirmatory use of the method would be for a researcher to have a strong cognitive model for the test such that specific OCSs may be predicted a priori. A strong cognitive model would allow a researcher to state in advance of data collection the distractor examinees would likely select given a specific amount of time to answer the item and a certain ability level. After data collection, the OCSs may be compared to what was expected a priori.

This method of including item response time in a distractor analysis has implication beyond construct representation. Namely, ability estimation is improved as information beyond correct and incorrect is included in estimation. Along these lines, Meyer (2006) discusses the implications of treating item response time as a nuisance variable. Another implication is that item response time must be measured to fully understand the process of cognition. Item responses alone only provide insight about the state of cognition. For the field of measurement to implement Embretson's (1999) cognitive design system framework, item response time will need to be routinely collected. As a result, computer based-testing may be the only way to gather the information necessary to implement more cognitively sophisticated measurement models.

Limitations

The analysis of the ILT was limited in a couple of ways. Although it was developed from a table of specification that included Bloom's taxonomy, no specific cognitive model was employed when developing the test. As such, the analysis was exploratory and was a first step at

developing a deeper understanding of examinee cognition. A second limitation was the nature of the ILT. An information literacy test, much like a reading or art test, does not lend itself to cognitive modeling as much as, say, a mathematics test. Identifying item content that is consistently associated with different OCSs can be difficult.

Limitations to using kernel smoothing as a way to include response time must also be mentioned. First, when used in an exploratory manner, additional research is needed to test hypotheses that may develop. The statements made about examinee cognition and the ILT item could really only be confirmed using think aloud protocols or other techniques. Nevertheless, rich information and plausible hypotheses may be obtained from the time-based distractor analysis. Second, parametric models that make specific assumptions about the speed-accuracy trade off function may be more effective techniques for the confirmatory approach mentioned earlier. Still other models may be better at such a confirmatory approach (see Wilson & De Boeck, 2005).

References

- Association of College and Research Libraries (ACRL; 2000). *Information literacy and competency standards for higher education*. Retrieved from <http://www.ala.org/ala/acrl/acrlstandards/informationliteracycompetency.htm> on March 30, 2005.
- Barton, K. E., & Huynh, H. (2003). Patterns of errors made by students with disabilities on a reading test with oral reading administration. *Educational and Psychological Measurement, 63*, 602-614
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.
- Dragow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143-165.
- Embretson, S. E. (1999). Cognitive psychology applied to testing. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.) *Handbook of Applied Cognition*. New York: John Wiley & Sons.
- Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Berlin: Springer-Verlag.
- Levine, M. V., & Drawsgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43*, 675-685.
- Meyer, J. P. (2006). *Kernel Estimation of Item Characteristic Surfaces that Incorporate Item Response Time*. Manuscript in preparation.
- R Development Core Team (2004). *R a language and environment for statistical computing*. Vienna, Austria: R Project Foundation. Retrieved from <http://www.R-project.org>.

- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611-630.
- Ramsay, J. O. (1997). A functional approach to modeling test data. In van der Linden, W. J., & Hambleton, R. K. (Eds.) *Handbook of modern item response theory*. New York: Springer.
- Roskam, E. E. (1997). Models for speed and time-limit tests. In van der Linden, W. J., & Hambleton, R. K. (Eds.) *Handbook of modern item response theory*. New York: Springer.
- Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213-232.
- Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. New York: John Wiley & Sons.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, *52*, 589-617.
- Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement*, *13*, 201-214.
- Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.) *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.) *Test theory for a new generation of tests*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In van der Linden, W. J., & Hambleton, R. K. (Eds.) *Handbook of modern item response theory*. New York: Springer.

Van Breukelen, G., J., P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*, 70, 1-18.

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. H. (1997). A logistic model for time limit tests. In van der Linden, W. J., & Hambleton, R. K. (Eds.) *Handbook of modern item response theory*. New York: Springer.

Wilson, M., & De Boeck, P. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18, 163-183.

Table 1

Item and Response Time Statistics

Item	\hat{p}	r_{pbis}	Response Time in Seconds		
			$\gamma(t_j)$	$IQR(t_j)$	$\hat{p}(item, \lambda_i)$
A	0.93	0.53	17	14	0.24
B	0.94	0.24	18	11	0.02
C	0.65	0.26	45	54	0.09
D	0.55	0.41	31	27.25	0.34

Figure 1. Option Characteristic Curves for Item A

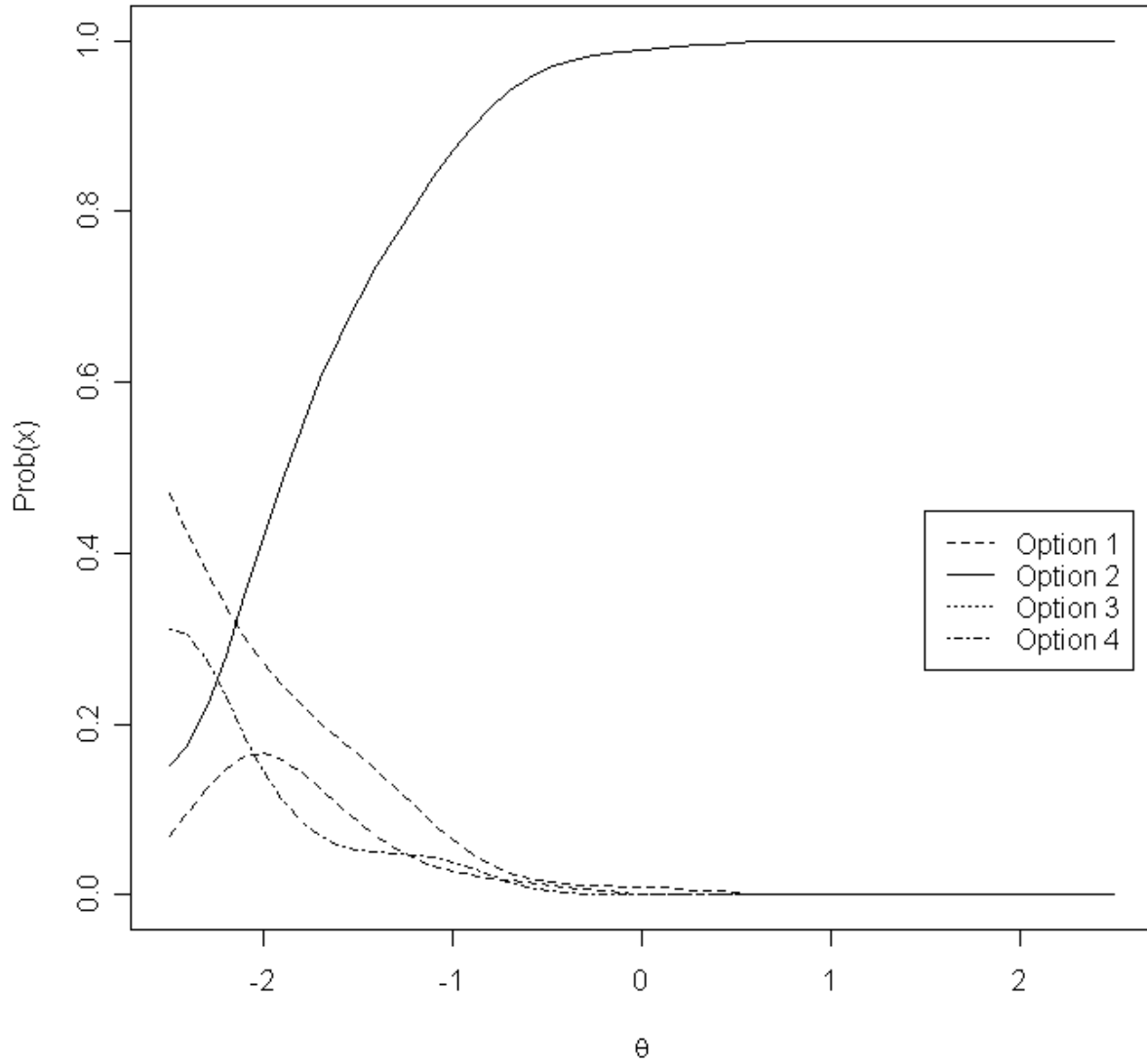


Figure 2. Option Characteristic Surfaces for Item A

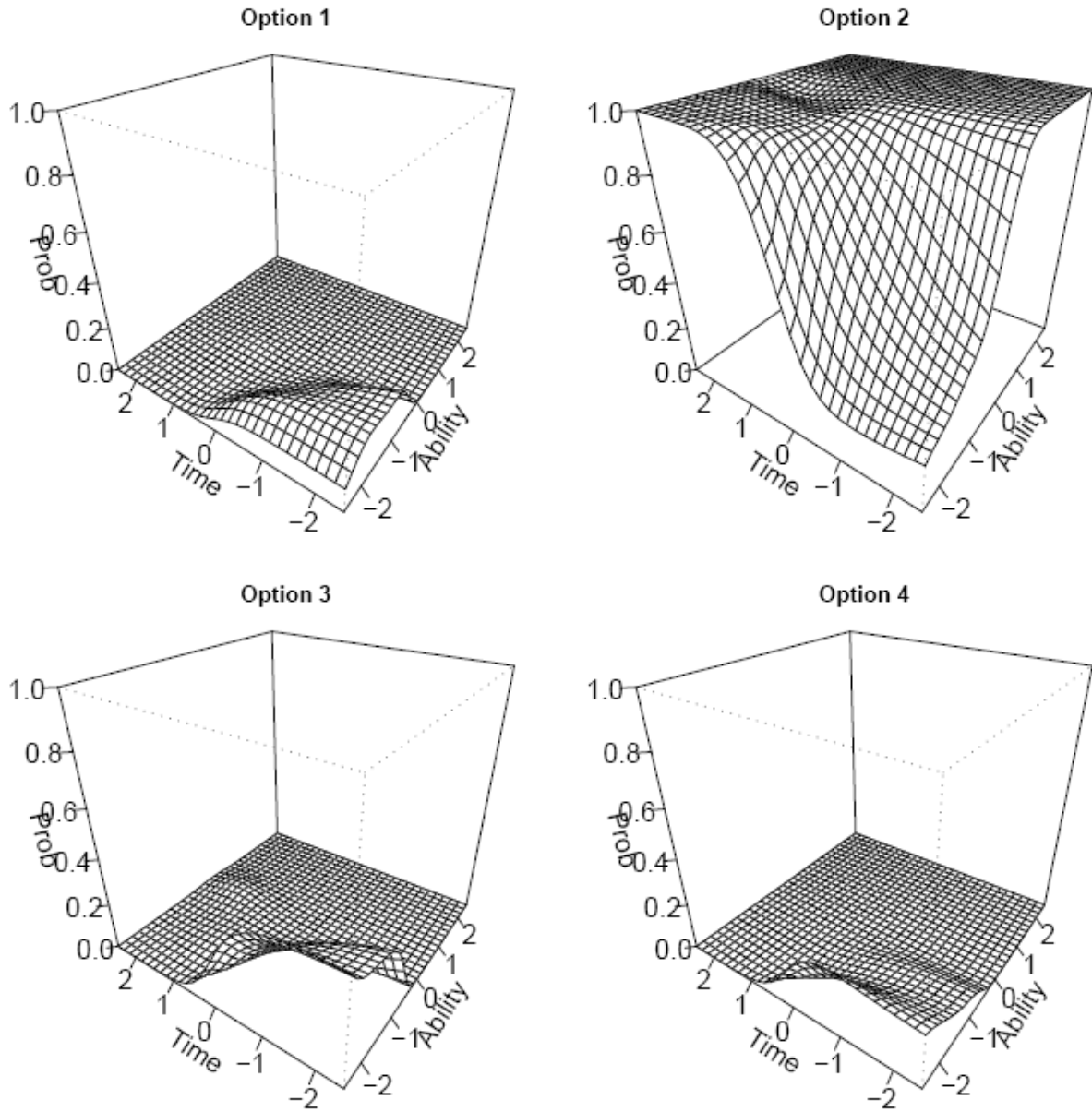


Figure 3. Option Characteristic Curves for Item B

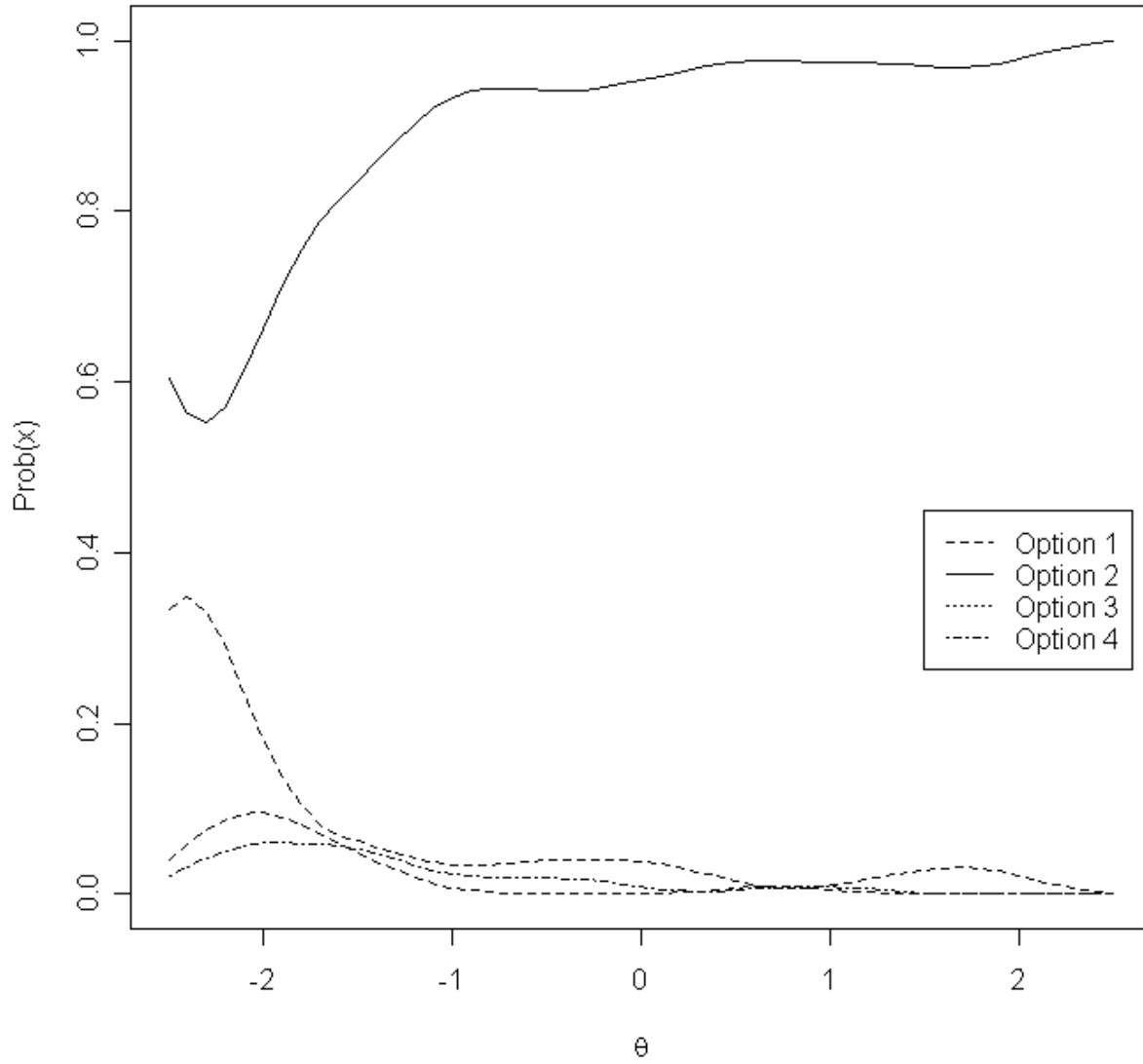


Figure 4. Option Characteristic Surfaces for Item B

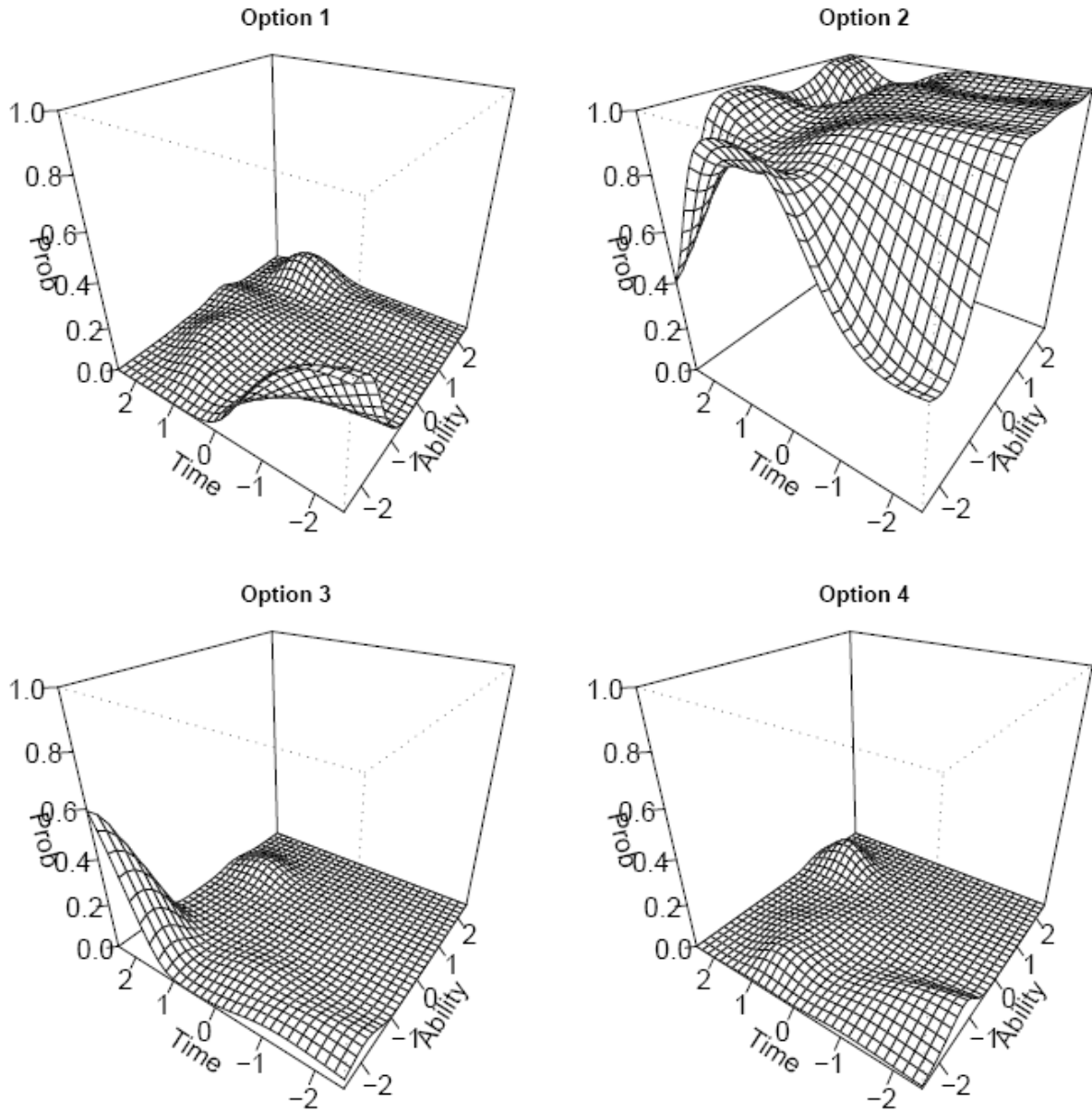


Figure 5. Option Characteristic Curves for Item C

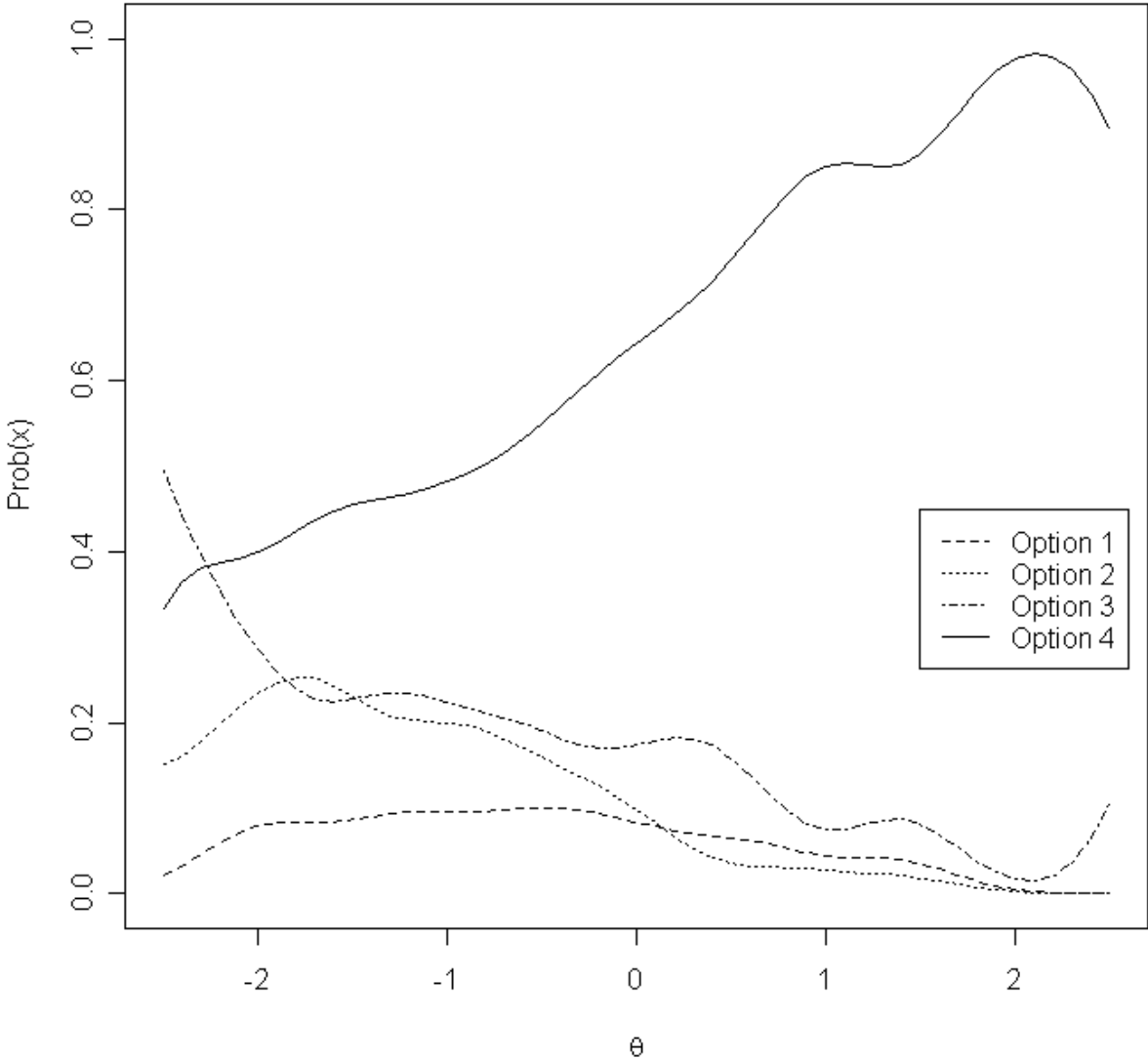


Figure 6. Option Characteristic Surfaces for Item C

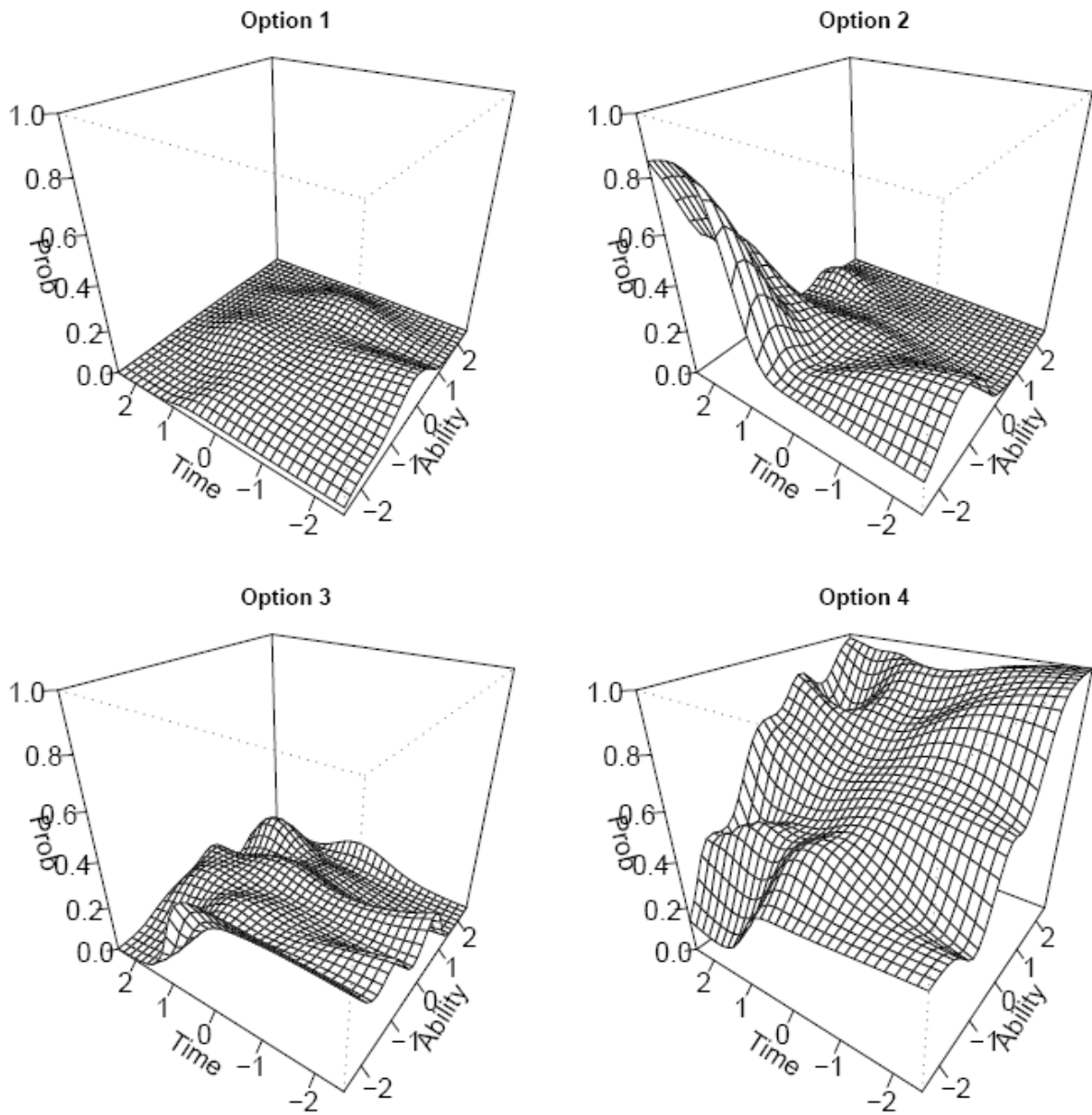


Figure 7. Option Characteristic Curves for Item D

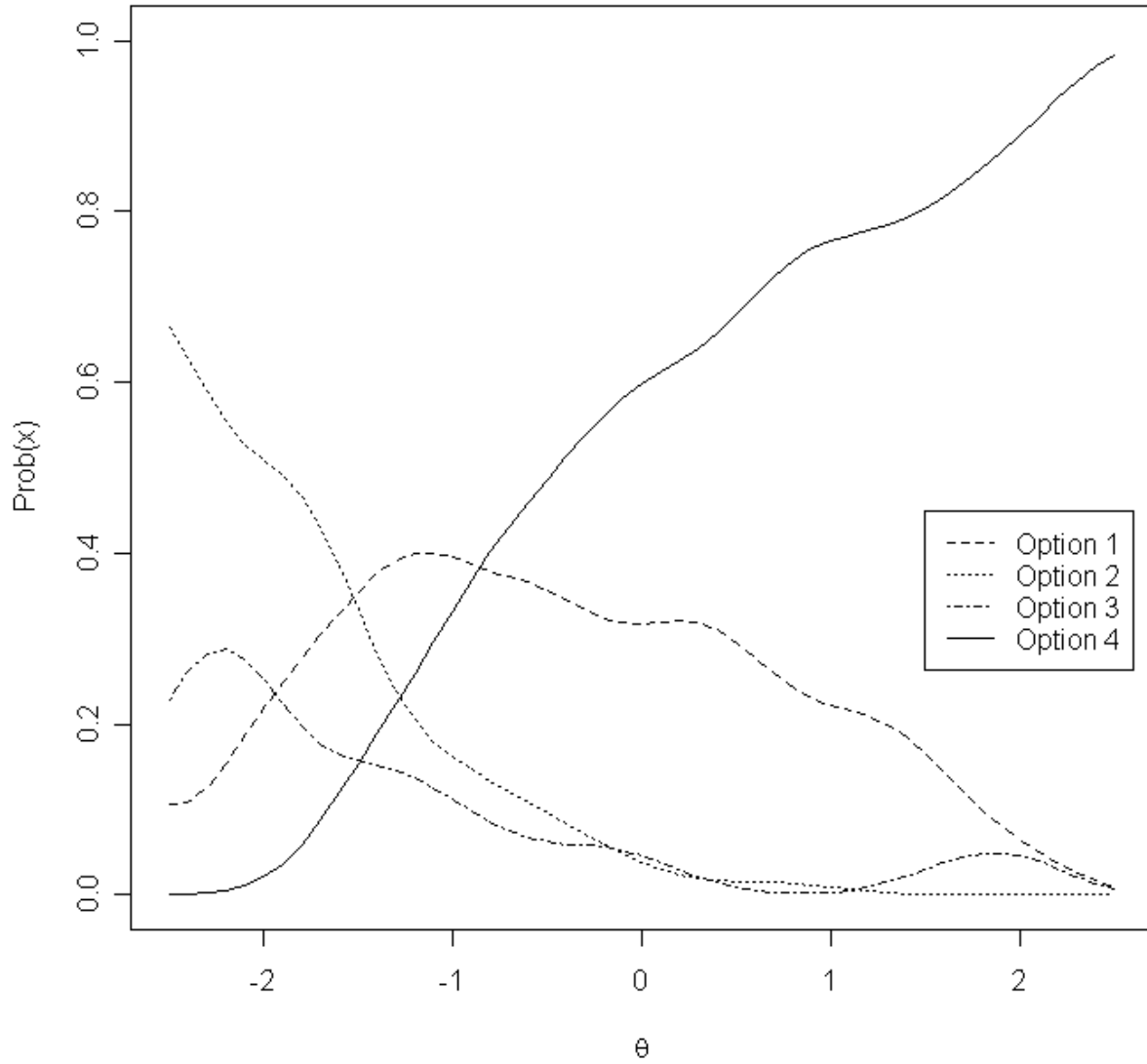


Figure 8. Option Characteristic Surfaces for Item D

