Behavior Research Methods
2009, 41 (2), 499-506
doi:10.3758/BRM.41.2.499

SIMPOLYCAT: An SAS program for conducting CAT simulation based on polytomous IRT models

Ssu-Kuang Chen
National Chiao Tung University, Taiwan, Republic of China

and

Karon F. Cook
University of Washington, Seattle, Washington
and Evanston Northwestern Healthcare, Evanston, Illinois

A real-data simulation of computerized adaptive testing (CAT) is an important step in real-life CAT applications. Such a simulation allows CAT developers to evaluate important features of the CAT system, such as item selection and stopping rules, before live testing. SIMPOLYCAT, an SAS macro program, was created by the authors to conduct real-data CAT simulations based on polytomous item response theory (IRT) models. In SIMPOLYCAT, item responses can be input from an external file or generated internally on the basis of item parameters provided by users. The program allows users to choose among methods of setting initial θ, approaches to item selection, trait estimators, CAT stopping criteria, polytomous IRT models, and other CAT parameters. In addition, CAT simulation results can be saved easily and used for further study. The purpose of this article is to introduce SIMPOLYCAT, briefly describe the program algorithm and parameters, and provide examples of CAT simulations, using generated and real data. Visual comparisons of the results obtained from the CAT simulations are presented.

Computerized adaptive testing (CAT) has been extensively studied and implemented in ability and achievement testing. Large-scale assessments such as the Graduate Record Examination (GRE) offer CAT as a test administration alternative to the traditional paper-and-pencil format. Although most CAT applications have used dichotomously scored items, the operational characteristics of CAT based on polytomous item response theory (IRT) have been studied systematically since the late 1980s. Dodd, De Ayala, and Koch (1995) evaluated and summarized item selection methods, stopping rules, and other CAT parameters in the context of polytomously scored items. A series of studies were conducted to compare ability/trait estimation methods in CAT based on polytomous IRT models (Chen, 1997; Chen, Hou, & Dodd, 1998; Chen, Hou, Fitzpatrick, & Dodd, 1997; Gorin, Dodd, Fitzpatrick, & Shieh, 2005; Penfield & Bergeron, 2005; Wang & Wang, 2001, 2002). Other issues pertinent to high-stakes testing have been examined, including item exposure rate and item security (L. L. Davis & Dodd, 2008; L. L. Davis, Pastor, Dodd, Chiang, & Fitzpatrick, 2003; Pastor, Dodd, & Chang, 2002).

Recently, there has been increased use of IRT in the measurement of health care outcomes. Typically, these outcomes are scaled using more than two response options. The testing efficiency afforded by CAT makes it particularly attractive in health outcomes research, where respondents often suffer from fatigue and other symptoms of compromised health and for whom completing long measures could be an undue burden. Polytomous CATs have been developed to assess a variety of health outcomes including headache impact (Ware et al., 2003), anxiety (Walter et al., 2007), fatigue (K. M. Davis, Lai, Hahn, & Cella, 2008), and physical function in children (Mulcahey, Haley, Duffy, Pengsheng, & Betz, 2008). The number of CAT-based health outcome assessments can be expected to grow substantially in coming years because of increased interest among health outcomes researchers and substantial federal funding earmarked for the development of such measures. For example, in 2004, the National Institutes of Health (NIH) funded the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative. This initiative established a national collaborative network to create a publicly available system for measuring patient-reported outcomes, using IRT applications including CAT (Cella, Gershon, Lai, & Choi, 2007; Cella, Yount, et al., 2007; DeWalt, Rothrock, Yount, & Stone, 2007; Reeve, Burke, et al., 2007; Reeve, Hays, et al., 2007; Rose, Bjorner, Becker, Fries, & Ware, 2008).

Because of the diversity of patient population and health problems, the measurement of health outcomes presents unique features and challenges (Thissen, Reeve, Bjorner,

S.-K. Chen, schen75025@gmail.com

& Chang, 2007). New options for CAT that have not received consideration in the educational context have been proposed for the measurement of health outcomes. For example, greater precision at clinically relevant ranges of the θ continuum can be obtained by setting the precision criterion for terminating CAT higher near the less healthy end of the θ continuum and lower at the more healthy end. Application of polytomous CAT for Likert-type scales, such as personality tests and attitude questionnaires, has drawn attention from researchers (MacDonald, 2003). For example, Hol and colleagues developed a CAT version of a polytomously scored motivation scale (Hol, Vorst, & Mellenbergh, 2007). Because item characteristics, assessment formats, item banking strategies, and other aspects of CAT can vary in different contexts, it is important to test features and options for CAT before its implementation. Particularly helpful would be research that evaluates CAT innovations suggested by new contexts.

A real-data simulation (i.e., post hoc simulation) of CAT is an important step in developing an operational CAT, since it allows CAT developers to evaluate various CAT testing parameters and methods prior to live testing (Weiss, 2005). SIMPOLYCAT is an SAS program that allows such simulations using polytomous IRT models. The fact that the SAS programming language is used should ensure that SIMPOLYCAT can be easily adopted by interested users. The program is designed to include basic options and features of CAT to meet the needs of most researchers. However, for proficient SAS users, it is possible to modify the program to meet special needs.

The purposes of this article are to (1) describe the SIMPOLYCAT program with a brief introduction of the CAT algorithm, data input requirements, output files, and features of the program, and (2) provide examples that demonstrate the operation of SIMPOLYCAT.

SIMPOLYCAT: An SAS Program

SIMPOLYCAT is appropriate for use with polytomously scored items such as Likert-type questions and short-answer questions for which partial credit can be awarded. The program contains an SAS macro and the invocation of the macro. The macro is called and processed according to the default or user-specified parameters. The user need not be an expert in the SAS macro language to run the program. However, basic knowledge regarding data input, format for data steps, macro invocation, and SAS program submission on the SAS system is required. Skills in SAS macro programming, such as macro variable naming, will allow more efficient use of the program. The program requires SAS Windows Version 6.07 or later with Base module installed.

In SIMPOLYCAT, item responses can be input from an external file or generated internally on the basis of item parameters provided by the user. The program combines item parameter data and item response data to create rows containing the variables representing item parameters and the variables representing item responses for each examinee. The variables representing item parameters are defined as elements of a two-way array in which the first element is item number and the second element is category number. Through the use of arrays and SAS DO-TO-END loops, a variety of calculations involved in CAT are conducted in an SAS data step. Macro statements such as %IF-%THEN, %DO-%END, and %ELSE-%DO-%END are used to conditionally select a routine to continue on the basis of the parameters specified in the invocation portion of the macro. The algorithms used in SIMPOLYCAT are based on that of Dodd, Koch, and De Ayala (1989), with revisions and the inclusion of additional functions.

CAT algorithm. In the beginning of the CAT process, for each examinee, the initial θ estimate is set. Item information is computed for each item in the item pool for the given level of θ, and then an item is selected from the pool to administer to the examinee. After the item is selected, responses to that item are identified for each person. The responses are either actual responses (input into the program) or responses simulated to fit the IRT model.

It is impossible to obtain a maximum likelihood estimate (MLE) for persons whose response to the first item is in the lowest or the highest category (Dodd et al., 1989). The issue persists until a nonextreme category response is obtained. Dodd et al. (1989) noted that the MLE $\hat{\theta}$ is unstable if the same category is endorsed for the first few items, and a variable step size procedure can be used to modify the most recent $\hat{\theta}$ until responses in two different categories occur. Since SIMPOLYCAT allows CAT simulation in which the number of item categories does not need to be the same for each item, the variable size procedure is employed only when extreme category responses are obtained for the first few items. In previous research, weighted likelihood estimation (WLE) failed to converge because of the extreme category responses in the beginning of CAT (Chen, 2007; Gorin et al., 2005). Thus, for MLE and WLE, SIMPOLYCAT employs a variable step size procedure to estimate θ until responses in nonextreme categories are obtained. With the variable step size procedure, for higher category responses, θ is estimated at halfway between the initial $\hat{\theta}$ and the value of the highest step difficulty in the item pool; for lower category responses, θ is estimated at halfway between the initial $\hat{\theta}$ and the value of the lowest step difficulty. After this initial estimate is obtained, test information and standard errors (*SE*s) of $\hat{\theta}$ are computed (*SE* is the inverse square root of test information). The CAT stopping criteria are evaluated to determine whether the CAT process will terminate or continue. If the CAT process continues, the items that have not been administered are again searched, using the method specified by the user. As before, the original response string for the examinee is checked for the actual (or simulated) category response for this item. The $\hat{\theta}$ and *SE* are then calculated on the basis of the provisional response string. The CAT process is repeated until the stopping criteria are satisfied.

Input parameters. The parameters used in the macro program are specified in the macro invocation portion of the SAS program. Some of the parameters are required, and no default values are preset. For the optional parameters, default values are provided by the program. Input parameters can be divided into three categories:

(1) model specification and data input, (2) CAT process, and (3) data output.

The program requires users to specify one of four available polytomous IRT models: the generalized partial credit model (Muraki, 1992), the partial credit model (Masters, 1982), the graded response model (Samejima, 1969), or the rating scale model (Andrich, 1978). Required input includes the item parameter data that are used for the item response generator or for the CAT run. The number of item step difficulties (or category boundaries) allowed in SIMPOLYCAT is one to nine. The number of item response categories is not required to be the same for each item, and there is no restriction on the number of items. If the response data were generated internally, the number of simulees needs to be provided and should be less than 10,000. In SIMPOLYCAT, there are no restrictions on the number of examinees when the item response data are provided from a file. However, if the number of examinees or items is extremely large, processing speed will be sacrificed.

Depending on the goals of the CAT simulation, the item parameters can be either calibrated or known. For a real-data CAT simulation, item parameters must be calibrated using software packages such as MULTILOG (Thissen, 2003) or PARSCALE (Muraki & Bock, 2003) and then input into SIMPOLYCAT. Users must specify the name and location of the item parameter data file, as well as that of all input and output files. Typically, the item response data are input from a file. However, if item responses are not available, random data can be generated on the basis of the provided item parameters and an assumed distribution (normal, uniform, or beta distribution). The generated item response data are saved in a text file that contains examinee ID, item responses, and the generated θ that can be used to compare with the estimated θ obtained from the CAT simulation.

The parameters for the operational procedures of CAT include (1) options for specifying initial θ, (2) options for calculating item information for item selection, (3) number of a set of most informative items from which an item is randomly selected for administration, (4) name of the data file that contains the item IDs of items that are locally dependent, (5) options for trait/ability estimation, (6) first n number of examinees in the item response data included in the CAT run, (7) maximum number of items administered in CAT, and (8) *SE* of θ estimate stopping criterion.

The macro program provides three options for specifying initial θ. First, the user can elect to set the initial θ to zero for every examinee. Second, initial θ values can be selected randomly from a range of θ values (e.g., -1 to 1); the lower and upper bounds of this range are entered in the macro invocation portion. The final option is to input initial θ values from a file. If the last choice is selected, the file name must be provided.

SIMPOLYCAT provides maximum posterior-weighted information (MPI) as a Bayesian approach to item selection, which is a promising alternative to maximum item information (MII) (Penfield, 2006). MPI is obtained by computing the expected information for the examinee by taking into account the examinee's posterior distribution of θ. If MPI is selected, the number of quadrature points and the prior must be specified. The number of items in the set of most informative items at the current estimated θ must be assigned. From this set, one item is randomly selected for administration in CAT. An SAS routine using this parameter is invoked to control item exposure rate. If the size is one, no item exposure control is invoked. Since content-balancing issues are less of a concern for Likert-type scales, content-balancing constraints are not applied in SIMPOLYCAT.

If one or more subsets of items in the item bank are locally dependent, a routine can be enabled to avoid selecting more than one item per locally dependent subset. Local dependence occurs when, after conditioning on θ, there remains an association among item responses. This may occur, for example, when items share very similar content. The name, location, and format of the file specifying the locally dependent items must be entered in the macro invocation statement.

The options for ability/trait estimation include MLE, expected a posteriori (EAP), maximum a posteriori (MAP), and WLE (Warm, 1989). The number of quadrature points for EAP estimation is 20 by default, but it can be changed by the user. A prior distribution (i.e., uniform or normal) is used for EAP and MAP estimations.

So that users can preevaluate the CAT simulation before including all examinees in a lengthy CAT routine, the program allows users to select the first n examinees from the item response data file. Stopping criteria include fixed-length and fixed-precision stopping rules. For a fixed-length CAT, the parameter value for the maximum number of items administered is set to the desired value, and the value for the *SE* stopping criterion is set to an extremely low value that would never be reached (e.g., -99). For a fixed-precision CAT, the maximum number of items administered should be set to the number of items in the item pool, and the value of *SE* stopping criterion should be set to the desired value. A combination of the two stopping criteria can be applied so that, when either of the two criteria is satisfied, the CAT process is terminated.

The final SAS data set can be saved permanently and/or exported to a text format data file. The name of the saved text file must be provided in the macro invocation. Users can elect to delete the final temporary SAS data sets after the CAT process is completed.

Output data files. SIMPOLYCAT creates several SAS data sets during processing. All but two are deleted at the completion of the simulation run. One of the two retained data sets contains item-level details of the CAT run, such as the item selected, response to the item, provisional ability/trait estimate, provisional test information, and so forth. The other retained data set provides summary results for each examinee, including (1) final ability/trait estimate, *SE*, and the number of items administered in CAT; and (2) final ability/trait estimate, *SE*, and the number of items administered in full-length testing. The SAS data sets can be saved permanently by using LIBNAME to designate a directory path. The program also allows item-level information and final summary results for each examinee to be saved to text file format.

Typically, item-level data can be used to investigate how $\hat{\theta}$s obtained after an additional item is administered converge to the final $\hat{\theta}$. To evaluate CAT performance, users can compare final estimated θ and *SE*s of $\hat{\theta}$s between CAT and full-length testing. One of the output text files contains summary CAT results and full-length results that can be imported into a statistical graphic tool, such as Microsoft Excel, to create a visual plot for comparisons. If known θ (or generated θ) is available, CAT-based estimated θ is compared with known θ to obtain an estimate of bias across the θ continuum.

To illustrate how to use SIMPOLYCAT, two example simulations are presented below.

Illustration of Using SIMPOLYCAT

Simulated data CAT simulation. In this example, a CAT simulation based on the generalized partial credit model (Muraki, 1992) was conducted. The item pool used in the simulation was adapted from Chen (2007). Four items were created with four item category boundaries each that make the items center on 0.0. A set of item discriminations was selected to represent low (a = 0.4), medium (a = 1.0), and high (a = 1.6) discriminations. The 4 sets of item category boundaries were crossed with the three levels of item discrimination so that there were 12 sets of item parameters. The item pool with 72 items was created by producing 6 items for each of the 12 sets of item parameters. Using the SIMPOLYCAT program, 6,000 responses to the 72 items were generated. θs for simulees were randomly generated from a uniform distribution within a θ range of −4 to 4, so that θ values were approximately evenly distributed along the θ continuum. This allowed the evaluation of the performance of the CAT for simulees with extreme θs. Ten replications of the response data were generated to account for sampling variation.

An initial $\hat{\theta}$ of 0.0 was used to select the first item (on the basis of maximum item information), and MLE was employed to compute ability/trait estimates in the CAT simulation. For examinees with responses in extreme categories, prior to MLE, a variable step size procedure was used to obtain a new $\hat{\theta}$ after the administration of the first or first few items according to the algorithm described above. CAT simulations were terminated when a maximum of 48 items was administered or a minimum *SE* of 0.24 was reached. The macro invocation statement used for this CAT simulation is presented in Figure 1. Parameters for this example are noted on the figure.

The CAT results were compared with results from full-length testing. Conditional plots were generated for the

```
%SIMPOLYCAT(
GPCM,
I72_M.IPM,
FILE_LOC=G:\SIMCAT07\GPCMDATA\,
PARINPUT=@1 ITEM $3.0 @6 A @11 B1  @17 B2 @23 B3 @29 B4 @35 NUMCAT,
R_DATA=2,
   NO_E=6000,
   KNOWNT=UNIFORM,
   SEED1=6734,
   SEED2=7867,
   R_OUTDATA=GPCMSIM.DAT,
INI_TSELECT=1,
RANDMSEL=1,
SELECTINFO=MII,
LIDFILE=NULL,
ESTIMATION=MLE,
NE=ALL,
MAXNI=48,
SESTOP=0.24,
EXPODATA=YES,
OUTFILE=GPCM11111C.TXT,
SAVE=YES,
REMOVE=NO);
run;
```

Figure 1. Macro invocation statements for simulated-data computerized adaptive testing simulation. R_DATA=2, item response data were generated from a uniform distribution with 6,000 stimulees and were saved in GPCMSIM.DAT; INI_TSELECT=1, initial $\hat{\theta}$ was set to 0.0; RANDMSEL=1, no item exposure control was invoked; SELECTINFO=MII, maximum item information was used; LIDFILE=NULL, no local item dependence among items; ESTIMATION=MLE, maximum likelihood estimation was used; NE=ALL, all the stimulees were included; MAXNI=48 and SESTOP=0.24, either a maximum of 48 items or a minimum standard error of 0.24 was satisfied; EXPODATA=YES, the final results were saved in text format; OUTFILE indicates the name; SAVE=YES and REMOVE=NO, SAS data sets were saved permanently and not removed.

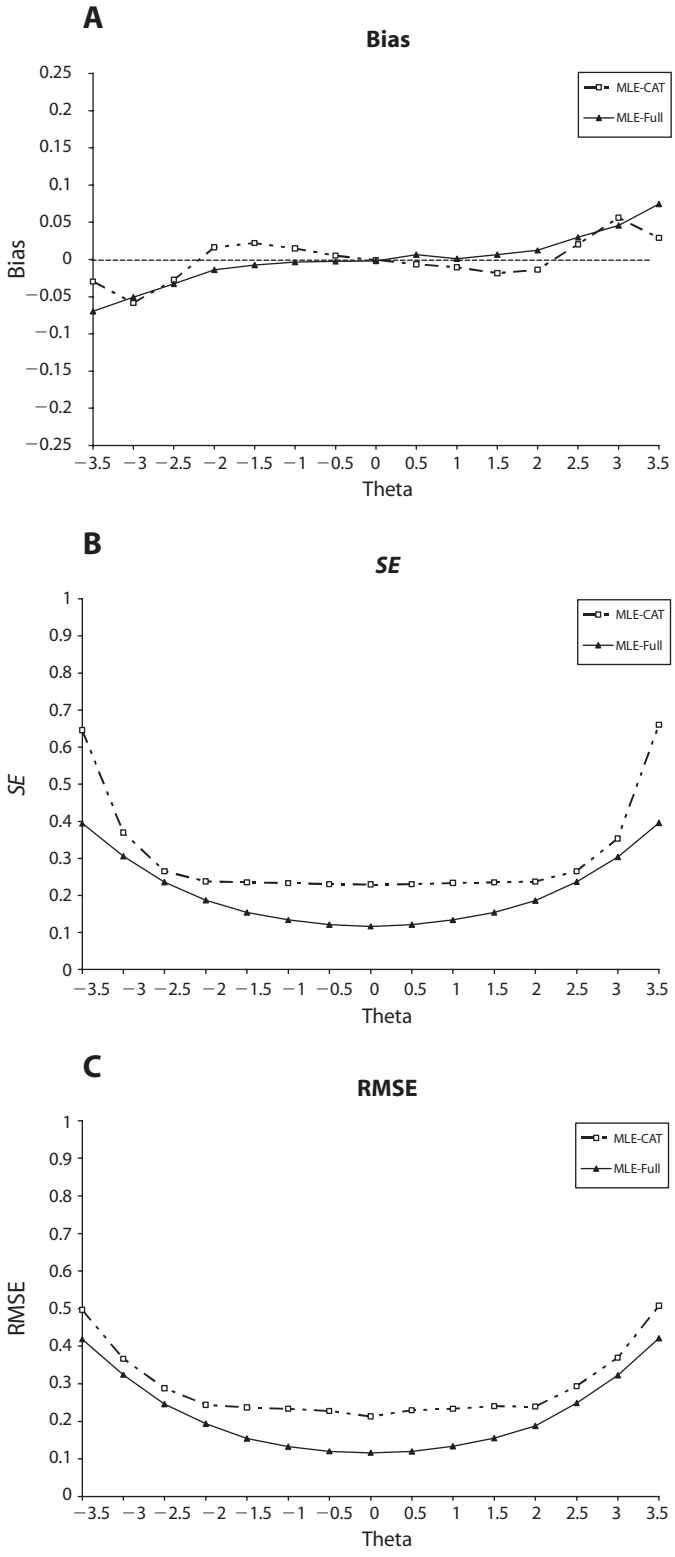

Figure 2. Conditional mean bias, standard error (*SE*), and root-mean squared error (RMSE) for the computerized adaptive testing (CAT) and full-length simulations. MLE, maximum likelihood estimation.

```
%SIMPOLYCAT(
GRM,
PS_PSC.PAR,
FILE_LOC=G:\SIMCAT07\,
PARINPUT=@1 ITEM $4.0 @6 A @16 B1 6.3 @26 B2 6.3 @36 B3 6.3 @46 NUMCAT 1.0,
R_DATA=1,
   R_FILE=PS_DATA.DAT,
   NIDCH=7,
INI_TSELECT=1,
RANDMSEL=1,
SELECTINFO=MII,
LIDFILE=NULL,
ESTIMATION=EAP,
   EAP_QUARPT=20,
   PRIOR=NORMAL,
NE=ALL,
MAXNI=30,
SESTOP=-99,
EXPODATA=YES,
OUTFILE=PS111121C2.TXT,
SAVE=YES,
REMOVE=NO);
RUN;
```

Figure 3. Macro invocation statements for real-data computerized adaptive testing simulation. R_DATA=1, item response data were input from PS_DATA.DAT; INI_TSELECT=1, initial $\hat{\theta}$ was set to 0.0; RANDMSEL=1, no item exposure control was invoked; SELECTINFO=MII, maximum item information was used; LIDFILE=NULL, no local item dependence among items; ESTIMATION=EAP, expected a posteriori estimation with 20 quadrature points and normal distribution was applied; MAXNI=30, fixed length of 30 items; EXPODATA=YES, final results were saved in text file format; OUTFILE indicates the name; SAVE=YES and REMOVE=NO, SAS data sets were saved permanently and not removed.

grand mean bias statistics, *SE*, and root-mean squared error (RMSE) on the basis of 15 intervals of θ (Figure 2). The conditional bias plot shows that, for both full-scale and CAT testing, MLE slightly overestimated ability/trait at higher levels of θ and underestimated ability/trait at lower levels of θ. The magnitude of the *SE* tended to be

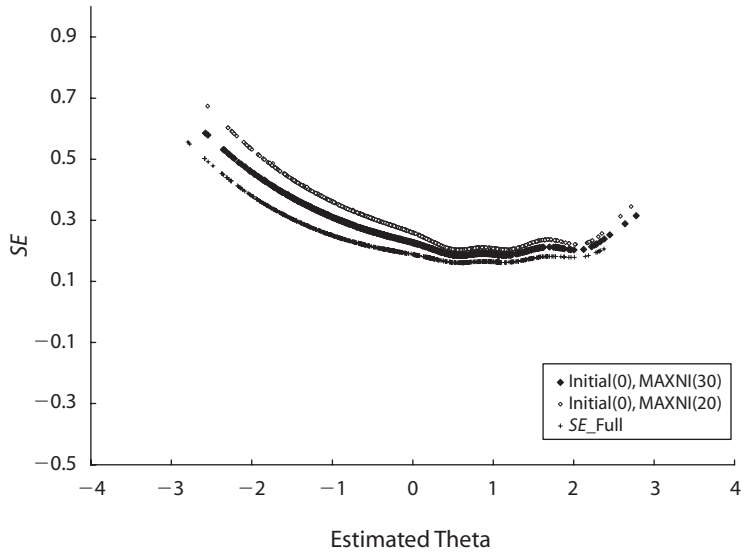

Figure 4. Scatterplots for standard errors (*SE*s) in computerized adaptive testing with different conditions.

higher at the extremes of the θ continuum. This finding is a function of the item pool used in the present study, since it provided less information at the extremes, as compared with the middle of the θ continuum. RMSE is an indicator of ability estimation accuracy or recovery. The values of the conditional RMSE statistic for the CAT were slightly lower in the center of the θ continuum.

Although the results above are based on only 10 replications, they concur with the results obtained in a generalized partial credit model-based simulation conducted by Wang and Wang (2002) for θ within the range of -3.5 to 3.5. This suggests that SIMPOLYCAT works properly and obtains results similar to those obtained with other CAT simulation tools.

Real-data CAT simulation. For this simulation, a real-data set was used that contained responses to 66 items of a scale measuring emotional distress. The items asked respondents about, for example, feelings of hopelessness, feeling frequently physically fatigued, and so forth. Respondents indicated their agreement or disagreement with the items by selecting one of the following options: *strongly disagree*, *disagree*, *agree*, or *strongly agree*. The data were provided by a researcher in Taiwan, with an agreement of use for program-testing purposes only and no content exposure.

The results of an item factor analysis of polychoric correlations indicated the existence of a single dominant factor. Items for which almost all responses were in the highest or lowest response categories were omitted from the analyses. Data from 714 respondents were used in the item calibration based on the graded response model (Samejima, 1969) using PARSCALE (Muraki & Bock, 2003). Three items that severely misfit the graded response model were deleted from the current simulation. The remaining 63 items were used in the CAT simulation. Since the purpose of this simulation was to illustrate the operation of SIMPOLYCAT, evaluation of results was limited to the impact of initial $\hat{\theta}$ setting (either 0.0 for all the respondents or randomly selected from a $\hat{\theta}$ range of -1 to 1) and number of items in a fixed-length CAT (20 or 30 items). The program used to simulate CAT with initial $\hat{\theta}$ equal to 0.0 and a fixed length of 30 items is presented in Figure 3. For all the conditions, EAP with 20 quadrature points and a normal prior were used.

Scatterplots were used to compare $\hat{\theta}$s in CAT with those obtained with full-length testing under different CAT conditions. With a 30-item CAT, the correlation between CAT-based $\hat{\theta}$s and full-length estimates was .99, slightly higher than that between estimates based on 20-item CAT and full-length estimates ($r = .98$). The correlations were the same in both initial $\hat{\theta}$ conditions (set to 0.0 or randomly selected from a $\hat{\theta}$ range of -1 to 1).

As was expected, the *SE*s obtained for the 20-item CAT were slightly higher than those obtained with the 30-item CAT. *SE*s for the 30-item CAT were close to those obtained in full-length testing (see Figure 4). It was also found that the discrepancies among *SE*s for the 20-item CAT, 30-item CAT, and full-length testing were larger for respondents near the lower end of the θ continuum (low in distress) than for those near the higher end of the θ continuum (high in distress). This was the result of the item bank's providing less test information near the lower end of the distress scale. The choice of initial $\hat{\theta}$ had little impact on trait/ability estimation in CAT.

In general, the results suggest that the 30-item CAT is a promising alternative to full-length testing when test precision near the high-distressed end is of clinical interest. If item security is of a concern, initial $\hat{\theta}$ randomly selected from a $\hat{\theta}$ range of -1 to 1 could be used.

Final Comments

Programs that simulate polytomous CAT may be developed in house by a variety of research institutes or testing companies, using different computer languages. The primary benefit of SIMPOLYCAT is that it runs in a familiar and popular computing environment and is free of charge. Portions of the program were adapted on the basis of a variety of systematic CAT studies, including evaluations of item security, content balance, and θ estimation in polytomous CAT (L. L. Davis et al., 2003; Gorin et al., 2005; Pastor et al., 2002). An early version of SIMPOLYCAT was used for an investigation of the effect of item bank on $\hat{\theta}s$, using various θ estimators and the graded response model (Chen, 2007). A CAT simulation using a cancer-related, health-related quality-of-life data set and a simulated CAT of the Modified Rolland–Morris Back Disability Questionnaire also used an early version of SIMPOLYCAT (Cook, Crane, & Amtmann, 2006; Cook et al., 2007). The present version of SIMPOLYCAT has many more options and should contribute to research on CAT based on polytomous IRT models.

Program Availability

The SIMPOLYCAT SAS program, its user guide, and examples of input and output files can be obtained by e-mailing Ssu-Kuang Chen at schen75025@gmail.com.

AUTHOR NOTE

The Patient-Reported Outcomes Measurement Information System (PROMIS) is a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring patient-reported outcomes in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS was funded by cooperative agreements to a statistical coordinating center (Evanston Northwestern Healthcare; PI, David Cella; U01AR52177) and six primary research sites: Duke University (PI, Kevin Weinfurt; U01AR52186), University of North Carolina (PI, Darren DeWalt; U01AR52181), University of Pittsburgh (PI, Paul A. Pilkonis; U01AR52155), Stanford University (PI, James Fries; U01AR52158), Stony Brook University (PI, Arthur Stone; U01AR52170), and University of Washington (PI, Dagmar Amtmann, U01AR52171). NIH Science Officers on this project are Deborah Ader, Susan Czajkowski, Lawrence Fine, Louis Quatrano, Bryce Reeve, William Riley, and Susana Serrate-Sztein. Development of SIMPOLYCAT was supported partially by a grant from the NIH (U01 AR 052177-01; PI, David Cella). The manuscript was reviewed by the PROMIS Publications Subcommittee prior to external peer review. See the Web site at www.nihpromis.org for additional information on the PROMIS cooperative group. The authors thank Barbara Dodd and Steve Fitzpatrick for their support in the early stage of the program development, as well as Hsin-Yi Chen for contributing the real data. The authors also thank four anonymous reviewers for their valuable comments. Correspondence concerning this article should be addressed to S.-K. Chen, Institute of Education, National Chiao Tung University, 1001 Ta-Hsueh Rd., Hsinchu 300, Taiwan, R.O.C. (e-mail: schen75025@gmail.com).

REFERENCES

ANDRICH, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, **43**, 561-573.

CELLA, D., GERSHON, R., LAI, J. S., & CHOI, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, **16**(Suppl.), 133-141.

CELLA, D., YOUNT, S., ROTHROCK, N., GERSHON, R., COOK, K., REEVE, B., ET AL. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care*, **45**(Suppl.), S3-S11.

CHEN, S.-K. (1997). A comparison of maximum likelihood estimation and expected a posteriori estimation in computerized adaptive testing using the generalized partial credit model (Doctoral dissertation, University of Texas at Austin, 1997). *Dissertation Abstracts International*, **58**, 453.

CHEN, S.-K. (2007). The comparison of maximum likelihood estimation and expected a posteriori in CAT using the graded response model. *Journal of Elementary Education*, **19**, 339-371.

CHEN, S.-K., HOU, L., & DODD, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational & Psychological Measurement*, **58**, 569-595.

CHEN, S.-K., HOU, L., FITZPATRICK, S. J., & DODD, B. G. (1997). The effect of population distribution and methods of theta estimation on CAT using the rating scale model. *Educational & Psychological Measurement*, **57**, 422-439.

COOK, K. F., CRANE, P., & AMTMANN, D. (2006, September). *Simulated CAT using the Modified Rolland–Morris Back Disability Questionnaire.* Paper presented at the NIH Patient Reported Outcome Measurement Information System national meeting, Bethesda, MD.

COOK, K. F., TEAL, C. R., BJORNER, J. B., CELLA, D., CHANG, C.-H., CRANE, P. K., ET AL. (2007). IRT health outcomes data analysis project: An overview and summary. *Quality of Life Research*, **16**(Suppl.), 121-132.

DAVIS, K. M., LAI, J.-S., HAHN, E. A., & CELLA, D. (2008). Conducting routine fatigue assessments for use in clinical oncology practice: Patient and provider perspectives. *Supportive Care in Cancer*, **16**, 379-386.

DAVIS, L. L., & DODD, B. G. (2008). Strategies for controlling item exposure in computerized adaptive testing with the partial credit model. *Journal of Applied Measurement*, **9**, 1-17.

DAVIS, L. L., PASTOR, D. A., DODD, B. G., CHIANG, C., & FITZPATRICK, S. J. (2003). An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. *Journal of Applied Measurement*, **4**, 24-42.

DEWALT, D. A., ROTHROCK, N., YOUNT, S., & STONE, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care*, **45**(Suppl.), 12-21.

DODD, B. G., DE AYALA, R. J., & KOCH, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, **19**, 5-22.

DODD, B. G., KOCH, W. R., & DE AYALA, R. J. (1989). Operational characteristics of adaptive testing procedure using the graded response model. *Applied Psychological Measurement*, **13**, 129-143.

GORIN, J. S., DODD, B. G., FITZPATRICK, S. J., & SHIEH, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement*, **29**, 433-456.

HOL, A. M., VORST, H. C. M., & MELLENBERGH, G. J. (2007). Computerized adaptive testing for polytomous motivation items: Administration mode effects and a comparison with short forms. *Applied Psychological Measurement*, **31**, 412-429.

MACDONALD, P. L. (2003). Computer-adaptive test for measuring personality factors using item response theory. *Dissertation Abstracts International*, **64**(2-B), 999. (University Microfilms No. AAINQ77103)

MASTERS, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149-174.

MULCAHEY, M. J., HALEY, S. M., DUFFY, T., PENGSHENG, N., & BETZ, R. R. (2008). Measuring physical functioning in children with spinal impairments with computerized adaptive testing. *Journal of Pediatric Orthopaedics*, **28**, 330-335.

MURAKI, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, **16**, 159-176.

MURAKI, E., & BOCK, R. D. (2003). Parscale for Windows (Version 4.0) [Computer software]. Lincolnwood, IL: Scientific Software International.

PASTOR, D. A., DODD, B. G., & CHANG, H. H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement*, **26**, 147-163.

PENFIELD, R. D. (2006). Applied Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education*, **19**, 1-20.

PENFIELD, R. D., & BERGERON, J. M. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Applied Psychological Measurement*, **29**, 218-233.

REEVE, B. B., BURKE, L. B., CHIANG, Y.-P., CLAUSER, S. B., COLPE, L. J., ELIAS, J. W., ET AL. (2007). Enhancing measurement in health outcomes research supported by agencies within the US Department of Health and Human Services. *Quality of Life Research*, **16**(Suppl.), 175-186.

REEVE, B. B., HAYS, R. D., BJORNER, J. B., COOK, K. F., CRANE, P. K., TERESI, J. A., ET AL. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, **45**(Suppl.), 22-31.

ROSE, M., BJORNER, J. B., BECKER, J., FRIES, J. F., & WARE, J. E. (2008). Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Clinical Epidemiology*, **61**, 17-33.

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

THISSEN, D. (2003). Multilog for Windows (Version 7.0) [Computer software]. Lincolnwood, IL: Scientific Software International.

THISSEN, D., REEVE, B. B., BJORNER, J. B., & CHANG, C.-H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research*, **16**(Suppl.), 109-119.

WALTER, O. B., BECKER, J., BJORNER, J. B., FLIEGE, H., KLAPP, B. F., & ROSE, M. (2007). Development and evaluation of a computer adaptive test for "anxiety" (Anxiety-CAT). *Quality of Life Research*, **16**(Suppl.), 143-155.

WANG, S., & WANG, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement*, **25**, 317-331.

WANG, S., & WANG, T. (2002). *Relative precision of ability estimation in polytomous CAT: A comparison under the generalized partial credit model and graded response model.* San Antonio, TX: Harcourt Educational Measurement. (ERIC Document Reproduction Service No. ED477926)

WARE, J. E., JR., KOSINSKI, M., BJORNER, J. B., BAYLISS, M. S., BATENHORST, A., DAHLÖF, C. G., ET AL. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research*, **12**, 935-952.

WARM, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, **54**, 427-450.

WEISS, D. J. (2005). *Manual for POSTSIM: Post hoc simulation of computerized adaptive testing* (Version 2.0). St. Paul, MN: Assessment Systems Corporation.

(Manuscript received April 16, 2008;
revision accepted for publication December 6, 2008.)