

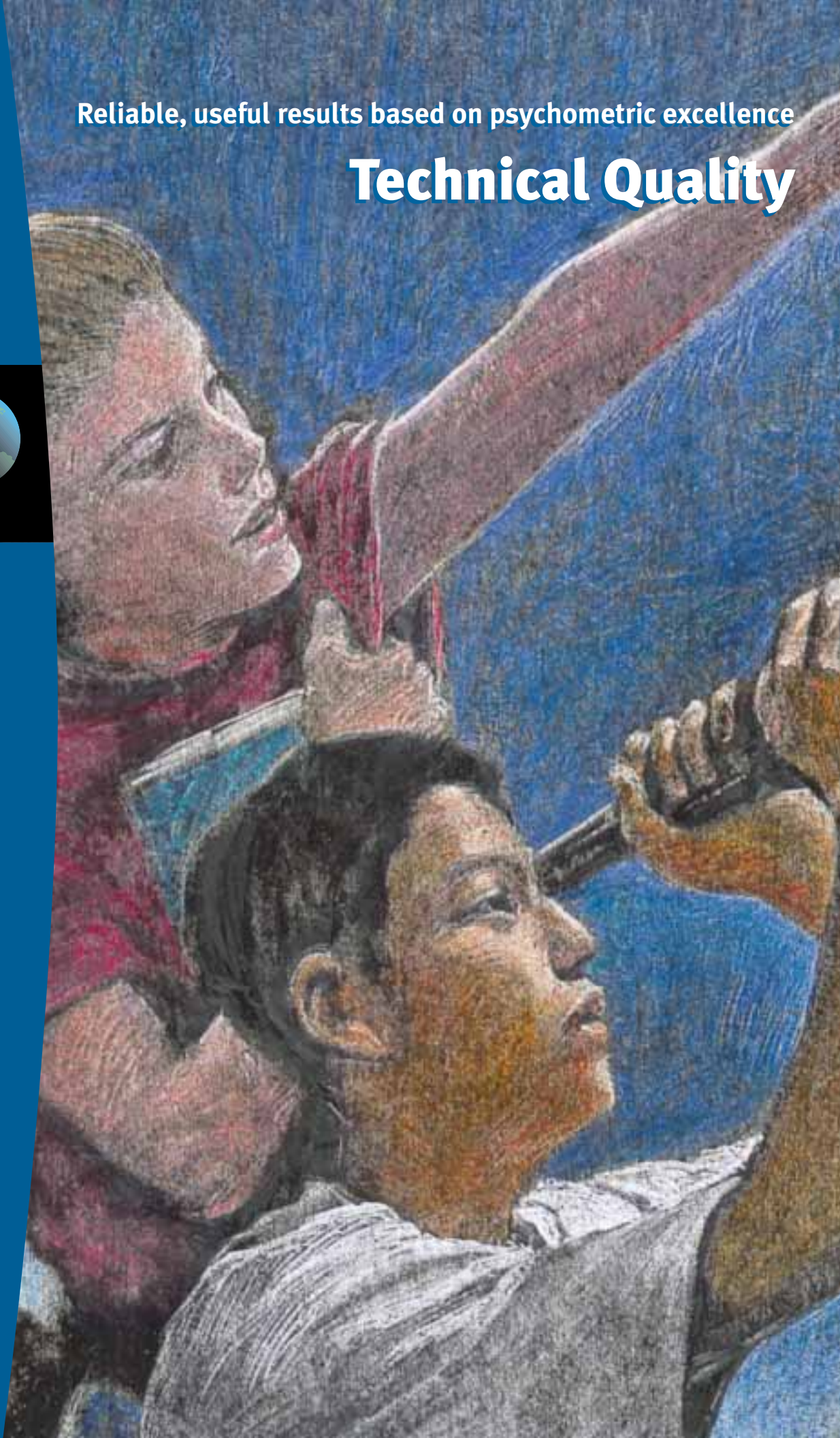
Reliable, useful results based on psychometric excellence

Technical Quality

TerraNova 

THE SECOND EDITION

 **CTB**
McGraw-Hill



EXCELLENCE IN ASSESSMENT:

Innovation combined with proven
research and development procedures.

Instructional **relevance**.

Advanced **measurement capabilities**.

This is

TerraNova, The Second Edition

—the newest member of the *TerraNova* family.

Important educational decisions require precise and accurate measurement information. The technical superiority of *TerraNova, The Second Edition* offers you significant advantages in today's challenging assessment environment.

Research Design for *TerraNova, The Second Edition*



The research design for the development of *TerraNova, The Second Edition* had two phases: tryout and national standardization.

- The tryout phase involved large samples of students across the country and provided empirical information about the quality of the tryout items. On the basis of these data, the best items were selected to appear in the final forms of the *TerraNova* CAT Multiple Assessments, Complete Battery, Survey, Plus tests, and the new *TerraNova* Algebra test.
- The standardization phase involved administration of these final forms to large, nationally representative samples of students and provided the normative data. These studies are described in the following sections.

For a more complete explanation of the research design, see the Technical Bulletins for *TerraNova, The Second Edition*.

Item Tryout

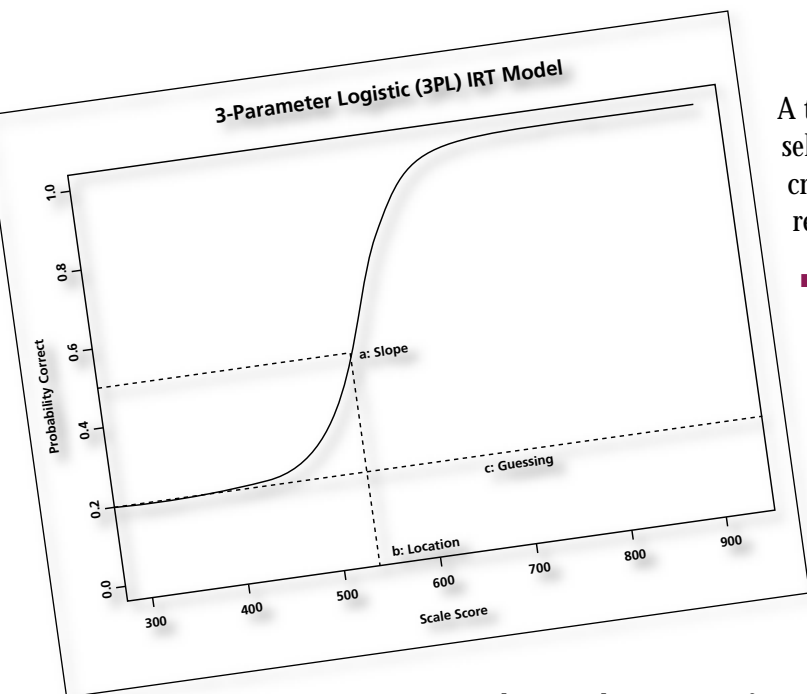
All of the items developed for the *TerraNova* CAT assessments were administered to samples of students in public and private schools across the country in the spring and fall of 1998. More than 100,000 students were involved in the tryout. In addition to the new items being tried out, each tryout book included anchor items selected from the corresponding test and level of *TerraNova* CTBS. These anchor items were used to link the *TerraNova* CAT tryout items to the *TerraNova* scale and to obtain estimates of national performance.



Tryout items were administered at several different grades. For example, items thought to be appropriate for Grade 5 students were tried out at Grades 4, 5, and 6. Gender and ethnicity of the participating students were also considered during this tryout phase to identify any potentially biased items. Empirical information about the quality of the items obtained in the tryout studies was used to select the best items for the final tests. Data obtained in the tryout studies were analyzed as described below.

Parameter Estimation

Assessments in the *TerraNova* family include both selected-response and constructed-response items. Because these types of items have different characteristics, two different Item Response Theory (IRT) models were used in analysis of the tryout data.



A three-parameter logistic (3PL) model was used for selected-response items. A two-parameter partial credit (2PPC) model was used for constructed-response items.

- The 3PL model defines each selected-response item in terms of three parameters: item difficulty or location, item discrimination, and the probability of a correct response by a very low-scoring student.
- The 2PPC model defines each constructed-response item in terms of two parameters: item discrimination, and a difficulty or location parameter for each score point assigned to the item.

For introductory discussions of IRT, see Bock (1977) or Chapter 11 in Allen & Yen (1979). For more advanced discussions of partial credit models, see Muraki (1990, 1992), Yen (1993), and Patz and Junker (1999).

CTB uses the 3PL and 2PPC models instead of the one-parameter, or Rasch model. Our extensive research has found systematic differences in the item discriminations of selected-response and constructed-response items. (See Fitzpatrick, Link, Yen, Burket, Ito, & Sykes, 1996.) The Rasch model, which assumes that all items are equally discriminating, cannot accommodate these differences. In contrast to the Rasch model, the IRT models that CTB uses are more flexible and more accurately represent student performance on items of different types.

The IRT models used in analysis of the tryout data were implemented using PARDUX, a PC-based parameter estimation software application developed at CTB (Burket, 1991). PARDUX estimates parameters simultaneously for right/wrong (dichotomous) and multipoint (polytomous) items using marginal maximum likelihood procedures implemented with the Expectation-Maximization (EM) algorithm (Bock & Aitkin, 1981; Thissen, 1982). Extensive simulation studies and comparisons have shown that PARDUX provides as precise or more precise parameter and ability estimates when compared with the results of other programs.

Differential Item Functioning Analysis

To minimize item bias, the tryout items were analyzed for differential item functioning (DIF). The degree to which items seemed to be functioning differently from the reference group was assessed for each focal group, using the Linn & Harnisch (1981) procedure implemented by PARDUX. Based on our analysis, each item was assigned a DIF rating. Item DIF ratings were used in the test construction process to produce final tests with average DIF ratings that were better than the average DIF rating for the entire pool of available items.

Item Selection

Item selection for this new edition of *TerraNova* was performed by content experts and reviewed by the psychometric staff at CTB. *TerraNova* CAT was designed to be an

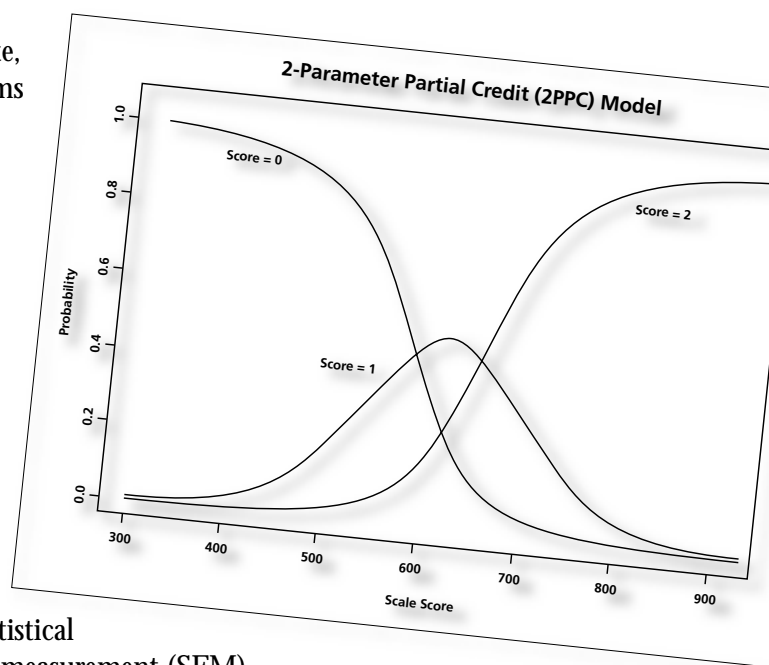
alternate form in the TerraNova family. That is, TerraNova CAT measures the same constructs in the same way that TerraNova CTBS does. As a result, what is learned about a student's knowledge, skills, and abilities would be the same if either TerraNova CAT or TerraNova CTBS were administered. Consequently, an important criterion of item selection was to meet the content specifications of TerraNova CTBS.

Developers selected items with statistical characteristics that were most similar to those of the items in TerraNova CTBS. To assist this process, the test developers used ITEMSYS, a PC-based software application developed at CTB (Burket, 1988), to select the final test items for our new TerraNova series.

ITEMSYS creates an interactive connection between the CTB developer and the test item database. The program monitors the impact of each decision made during the item selection process and offers a variety of options for grouping, classifying, sorting, and ranking items by highlighting key information as it is needed. (See Green, Yen, & Burket, 1989.)

ITEMSYS-based item selection consists of four tasks:

- **Initial screening.** The test developer selects a working item pool of manageable size from the larger tryout pool of items. ITEMSYS allows the test developer to track a variety of detailed information about each item, including the content objective to which the item is assigned, a descriptive phrase about the item, the association of the item with a passage or stimulus, a DIF rating, the item parameters, and a fit rating that indicates how well the data for this item fit the IRT model. The developer screens this pool for appropriateness to the target grade range and eliminates items that are clearly inappropriate.
- **Test selection.** Once this initial screening is complete, the developer uses the working pool to select the items for the test. Typically, the developer begins by specifying the number of items to be included in the test and the required number of items for each content objective. Based on these specifications, ITEMSYS automatically selects the best possible statistical combination of items for the test. These automatic selections can then be used as a reference set to which subsequent selections can be compared.
- **Statistical analysis.** Next, the developer examines the selected test to identify problems, including whether the test is too easy or too difficult for the target grade, contains DIF items, or does not match a parallel test form. To facilitate this process, ITEMSYS provides detailed graphic and statistical summaries on DIF, model data fit, standard error of measurement (SEM), and other test characteristics.
- **Iterative review.** A test developer works in collaboration with the research staff to ensure that the best items are selected. The flexibility and utility of the ITEMSYS program encourage multiple reviews to fine-tune the selection.





References

- Allen, M. J., & Yen, W. M. (1979). **Introduction to measurement theory**. Monterey, CA: Brooks/Cole.
- Bock, E. Darrell (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, 4, 21–33.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 66, 443–459.
- Burket, G. R. (1988). ITEMSYS [Computer program]. Unpublished.
- Burket, G. R. (1991). PARDUX [Computer program]. Unpublished.
- Fitzpatrick, A. R., Link, V., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. *Journal of Educational Measurement*.
- Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. *Applied Measurement in Education*, 2, 297–312.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons.
- Jolly, S. J., Johnson, R., Jones, B., & Abalus, J. (1986, April). **The effect of test speededness and random guessing on the validity of reading comprehension scores**. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.
- Lewis, D. M., Mitzel, H. C., Green, D. R. (1996). Standard Setting: A Bookmark Approach. In D. R. Green (Chair), *IRT-Based Standard-Setting Procedures Utilizing Behavioral Anchoring*. Symposium presented at the 1996 Council of Chief State School Officers 1996 National Conference on Large Scale Assessment, Phoenix, AZ.
- Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (April, 1998). **The Bookmark Standard Setting Procedure: Methodology and Recent Implementations**. Paper presented at the 1998 National Council for Measurement in Education annual meeting, San Diego, CA.
- Lin, M. H. (1986, April). **The impact of time limits on test behaviors**. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.
- Linn, R. L., & Harnisch, D. (1981). Interactions between item content and group membership in achievement test items. *Journal of Educational Measurement*, 18, 109–118.
- Macmillan/McGraw-Hill. (1993). **Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals**. New York, NY.
- McGraw-Hill. (1983). **Guidelines for Bias-free Publishing**. Monterey, CA.
- Munger, G. F., & Loyd, B. H. (1991). Effect of speededness on test performance of handicapped and nonhandicapped examinees. *Journal of Educational Research*, 85 (1), 53–57.
- Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14, 59–71.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Patz, R. J., and Junker, B. W. (1999). Applications and extensions of MCMC for IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24, 342–366.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.
- Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47, 175–186.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–213.



The
Measure | **of**
Success

CTB/McGraw-Hill

A Division of The McGraw-Hill Companies



20 Ryan Ranch Road
Monterey, California 93940-5703
800/538-9547
www.ctb.com

CTB/McGraw-Hill Regional Offices

Western Regional Office

20 Ryan Ranch Road
Monterey, California 93940-5703
831/393-7780

Northern Regional Office

12412 Powerscourt Drive, Suite 110
St. Louis, Missouri 63131-9998
314/821-1700

Southern Regional Office

3260 Peachtree Industrial Boulevard, Suite 20
Duluth, Georgia 30096-2547
770/622-4300

Published by CTB/McGraw-Hill, a division of The McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2001 by CTB/McGraw-Hill. All rights reserved. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

The materials shown in this publication are non-secure samples of secure test materials, some of which may be in unpublished form and under development by CTB. The final published version may have modifications.

California Achievement Tests, CTBS, SUPERA, and *TerraNova* are registered trademarks, and CAT and Classroom Connections are trademarks, of The McGraw-Hill Companies, Inc.