JACK W. McLAUGHLIN
Superintendent of Public Instruction

KEITH W. RHEAULT
Deputy Superintendent
Instructional, Research and Evaluative Services

DOUGLAS C. THUNDER
Deputy Superintendent
Administrative and Fiscal Services

STATE OF NEVADA

DEPARTMENT OF EDUCATION
700 E. Fifth Street
Carson City, Nevada 89701-5096
(775) 687-9200 • Fax: (775) 687-9101

SOUTHERN NEVADA OFFICE
1820 E. Sahara, Suite 205
Las Vegas, Nevada 89104-3746

(702) 486-6455
Fax: (702) 486-6450

January 27, 2003

MEMORANDUM

To: State Board of Education

From: Paul M. La Marca, Ph.D., Director
Richard Vineyard, Ph.D., Asst. Director

Re: Harcourt Resolution

We would like to update you on progress since the December State Board of Education meeting with respect to the Harcourt resolution. First, the independent quality assurance study was completed and we have attached the report summarizing findings and recommendations. Harcourt has agreed to implement all recommended changes, with the exception of certain psychometric services, at no additional cost to the state. In general, the review was favorable of Harcourt procedures. No significant deficiencies were noted.

Second, ongoing negotiations with Harcourt on potential services have occurred and have resulted in the following NDE recommended services: a) practices tests b) indicator reports c) pre-ID labeling, and d) on-site psychometric services. All recommended services help to strengthen the overall quality of the comprehensive Nevada assessment program. Pre-ID labels and on-site psychometric services are closely related to the accurate processing and scoring of student tests and are key additions to the program. On the following pages we have laid out a 4-year budget describing the receipt and cost of service.

The Department recommends that the State Board Act to accept the recommended services. Any budgetary change or service change will be brought forward to the State Board.

EXHIBIT AA Education Document consists of 11 pages
☑ Entire document provided.
☐ Due to size limitations, pages ____ through ____ provided.
A copy of the complete document is available through the Research Library (775-684-6827 or e-mail library@lcb.state.nv.us)
Meeting Date 2-24-04

An Equal Opportunity Agency

Practice Tests Grades 3 - 8	One-time cost -- $210,000 (14 tests - $15,000 per test) 6 – Reading, grades 3-8 6 – Math, grades 3-8 2 – Science, Grades 5 & 8	To include aligned practice tests per grade, per subject (aligned to content and cognitive domain specifications). Each test would include approximately 38 multiple choice items and 2 constructed response items. Camera ready files for the web by 9/1/04 (grades 3, 5, 8) or 1/1/05 (grades 4, 6, 7).
Content indicator Analysis report	2004-05 -- $30,000 2005-06 -- $20,000 2006-07 -- $20,000	This report would provide raw score performance on prioritized content indicators measured by at least 4 items. Paper reports to be provided by school, district, and state (one copy each) in grades 3, 5, and 8 for the 2004-05 school year (grades 3-8 thereafter).
On-Site Psychometrician	Per day -- $650.00	This includes salary and benefits
Pre-Id Labels	.20 per label	This is well below catalog price. Catalog is .46 per label and includes a setup fee per district of $360.00.

2003-04 School Year

Practice activities (all grades/all subjects)	$210,000
Psychometrics (15 days @ $650.00)	$9,750
Pre-Id Labels @ 20 cents per student & grade	----------
Total	$219,750

2004-05 School Year

Psychometrics (30 days @ $650.00)	$19,500
Pre-Id labels	$42,000
Item reports	$30,000
Total	$91,500

2005-06 School Year

Psychometrics (20 days @ $650.00)	$13,000
Pre-Id labels	$42,000
Item reports	$20,000
Total	$75,000

2006-07 School Year

Psychometrics (10 days @ $650.00)	$6,500
Pre-Id labels	$42,000
Item reports	$20,000
Total	$68,500

Grand Total	$454,750

Based on the total starting point of $485,000 and a to date expenditure of $36,000, there is approximately $449,000. Using funding already available from the previous Harcourt settlement, this shortfall can be made up.

Cc: Ada Woodward-Aguilar, Jack McLaughlin, Keith Rheault

Site Review of Harcourt's Quality Control Procedures

Richard Hill
The National Center for the Improvement of Educational Assessment, Inc.

January 14, 2004

Background

Harcourt Educational Measurement (Harcourt) is the contractor for three assessment programs in Nevada:

1. The Nevada 1994 High School Proficiency Exam, which tests adults in reading and mathematics. This test is a high school graduation requirement for adults who did not meet these requirements while in high school. The test is administered five times each year; over 2,500 people were tested during the 2002-03 school year.
2. The Nevada 1998 High School Proficiency Exam, which replace the earlier program. Students start taking the test in grade 10; they must pass it to get a high school diploma. One new form is administered each April; there are four other administrations each year using a previously used form.
3. The Nevada CRT program, which provides for testing of reading and mathematics in grades 3, 5 and 8. Science will be added to the program beginning with the next school year.

In the past two years, Harcourt staff made two production errors that led to incorrect reports being produced.

1. In 2002, the wrong passing score was set for the mathematics test on the 1998 HSPE. The passing score was set at a raw score of 42 when it should have been 41. As a result, 736 students who actually passed the test were reported to have failed.
2. In 2003, an incorrect raw score to scaled score conversion table was developed for the CRT program. As a result, virtually all students received an incorrect scaled score and about one-third were reported at the wrong achievement standard.

Because of these errors, Harcourt, through the request of the Nevada Department of Education, hired the Center for Assessment to create a team to review the quality control procedures used by Harcourt and make suggestions for the improvement of them. The Center selected a team of three people who had background in statewide assessment:

1. Richard Hill, the Executive Director of the Center and former president of Advanced Systems
2. Mark Moody, former director of assessment for Maryland
3. Pat DeVito, former director of assessment for Rhode Island

In addition, Harcourt arranged for Huixing Tang, President of eMetric, LLC, to join the review team. Short resumes of these four people are included as an appendix to this report.

The team made a site visit to Harcourt's corporate headquarters in San Antonio on December 15 and 16, 2003. Because of bad weather, Pat DeVito could not make the trip, but still reviewed

materials prior to the meeting. Richard Hill, Mark Moody and Huixing Tang spent a day and a half on site observing Harcourt's procedures and discussing them with Harcourt staff. All four consultants have reviewed and provided input to this report.

The Meeting

The meeting on December 15 began with introductions and an overview of the intent of the meeting. In attendance from Harcourt were:

1. Jack Dilworth, Chairman
2. Tom Rice, Senior Vice President, Operations
3. Jean Shimko, Vice President, Quality Assurance
4. John Olson, Vice President, Psychometrics and Research Services
5. Rudy Regalado, General Manager, Scoring Operations
6. Brandon Burgess, Director, Scoring Operations
7. Michael Young, Director, Psychometrics and Research Services
8. Ada Woodward, Program Manager
9. Zarko Vukmirovic, Senior Psychometrician
10. Jeff Davis, Psychometrician
11. William Garza, Senior Quality Assurance Coordinator
12. James Rodriguez, Quality Assurance Programmer Analyst
13. Ken Stallman, Senior Manager, Scoring Operations Planning and Analysis
14. Joyce McDonald, Director PASC (Scoring of constructed-response questions)
15. Roz Granderson, Coordinator, Program Management
16. Duncan MacQuarrie, National Measurement Consultant
17. Max Tudor, Measurement Consultant

Also in attendance was Richard Vineyard from the Nevada Department of Education. In addition, Jeff Galt, President and CEO, joined the meeting on the second day.

Following introductions and the overview, the group took a tour of the shipping, receiving, scanning, editing and scoring facilities. Following that, we returned to the conference room where we spent the remainder of the day and the following morning discussing Harcourt's quality assurance procedures.

Findings

Perhaps it is best to start with a summary. The panel found Harcourt's current quality control procedures to be of high quality. Some of the procedures being employed are improvements over their past practices, so we believe their quality control likely is significantly improved over what it was even a short time ago. While the panel has some recommendations for additional improvements that will be discussed later in this section, we were impressed with the procedures already in place. It is worth noting that no one on the panel had inside knowledge of the current quality control procedures of other contractors, so we cannot comment on whether other contractors' procedures are better or worse. Our judgment therefore is absolute, not relative: We find Harcourt's quality control procedures to be impressive and consistent with our best understanding of contemporary practices.

The tour of the facilities was impressive. The warehouse section of the building is large (225,000 square feet), well lit and ventilated, well organized and well maintained. Work areas are clearly delineated, have sufficient space, and clutter is kept to an absolute minimum. While our tour occurred during a slow period for processing, it was clear that the system has capacity for a substantial amount of materials. During this past May, 2.8 million student booklets were processed.

There are a series of controls that govern the processing of materials. Test booklets are sent through a machine that prints a unique number on each before shipment. That number provides a means of tracking whether all booklets have been returned, and if not, from which district they are missing. When materials are returned, they are logged in and discrepancies resolved. The materials from each shipment are placed on a separate cart, which carries them through the entire receiving process. Materials from different states are received at the same time. Each is placed on a different cart, and the carts delivered to sections in the facility that have responsibility for that particular contract. The implementation of a pre-printing procedure, whereby student identification and demographic information is printed on the booklets and answer documents prior to distribution, not only will facilitate materials tracking but should also improve the quality of the data by reducing the number of bubbling errors made by students and test administrators. The pre-printing procedure is scheduled for implementation with the 2004 test editions.

A series of quality assurance steps are employed to ensure error-free production. A control document called "Testmap" is created; Testmap becomes the source document for all information related to the test. Obviously, it is of paramount importance that Testmap is accurate. The panel repeatedly probed to find ways that Testmap could include an error that would affect the accuracy of tests and reports, but was left with the belief that Harcourt has thought through the process well. Each test question is assigned a company identification number (a "CID"); all information about the question is contained in a master file. At the beginning of the test development process, the items to be included are chosen, and Measurement Services creates an Excel file containing the CIDs for all the items. At least three times during the process of creating the tests, the items actually on the test are compared to printouts of the items drawn from the master file. There is a process of controls and sign-offs, ending with the Quality Assurance staff, to ensure that the questions on the actual test booklets, after printing, are the same as the questions with those same CIDs in the master bank.

Harcourt is getting more effective at using Testmap as a central quality control document. One particular example is the use of Testmap in the production of JCL to create standardized names for data files related to the processing of the test results. For example, the WINSTEPS input data file for the spring administration of eighth grade reading tests for Nevada this year would be name c:\PROJECTS\STNV3\SPR\MA08V1.DAT. Each character in that file name is dictated by the conventions Harcourt has adopted. Only a file with that name can be used as the input for WINSTEPS for that particular grade and subject combination for Nevada. A file with that name can only be created from data that has followed Harcourt's raw data processing system and has been signed off by a member of the Quality Assurance team. Thus, when WINSTEPS is run, there is assurance that the output relates to the proper grade and subject.

Another example of quality control is the use of a "test deck." When the answer sheets are printed, a sample of them are bubbled. Across the set, every bubble and every possible outcome is coded (for example, at least one answer sheet is bubbled in with the correct answer to every question; another has the wrong answer to every question), including every possible edit check. The coding is done according to a specification sheet that has been designed by Harcourt using input provided by the Department of Education. The test deck is then boxed and taken to the receiving department,

where the test deck is logged in as though it were materials being received from a district. The test deck is processed through scanning and editing, and the output checked against the specifications. Scanning and PASC (hand-scoring) scoring can begin on materials received from the districts only when a review shows that the test deck scan output file and the PASC data are correct and a sign-off is provided. Similarly, editing of live documents requires satisfactory editing of the test deck file and a sign-off on that procedure.

The errors that Harcourt made on Nevada's reports have not come from this section of processing. So far as the panel is aware, the raw data files have been produced with no or minimal errors. Thus, the systems that Harcourt have in place through the production of the raw data file appear to have worked well. The errors have come after the raw data file was produced but before the production of reports: that is, within the Psychometrics department. Therefore, the panel spent considerable time looking at the processes used there and the potential for error within that group.

Harcourt has recently hired new senior staff to manage the Psychometrics group and added new positions within that group to increase its capacity. The new management has initiated the standardization of naming conventions for data files (outlined above) and a process for comparing the results of the new year's data to those of the previous year. Both are commendable improvements that should reduce the probability of error within that group. The panel believes that implementation of the following recommendations would reduce the probability of error even further:

1. **Specify a list of items to be checked, the criterion for each item that will trigger a cautionary flag, and maintain a log of the resolution of all flags**. Harcourt staff identified a series of items that staff looked at to gain a sense of whether data might be in error. For example, if the difficulty of an item was extreme, that might be an indicator that the wrong key was used to score it; or if the count of the number of students in the final data file was considerably different from the number of students that were supposed to be tested, that would be an indicator that the file was incomplete. However, there was not a complete list of all the checks that were made, and for some of the checks that were listed, the criteria that would trigger a second look at the data was not specified (for example, one of the checks is to determine whether a "high volume" of students have been designated as being in special education; "high volume" should be operationally defined). For the processes for which such a list and criteria existed, there was no log maintained of the occurrences and the resolution of them. Often, it was left to the judgment of the person looking at the data to determine what should be looked at and what the criteria were that would raise a concern. As a result, error-checking is somewhat idiosyncratic to the person doing the checking, and the experience and skills that individual might have.

   The panel recommends that there be a standardized list of the items to be checked throughout the data flow. Each item on the list should have a trigger value; any observed result outside the that amount should precipitate further investigation as to whether the data are in error or the result is simply an unusual event; and a log should be maintained of the occasions when these events occur and what the resolution was. Once these checks have been agreed upon, most of them could be programmed and done automatically. For example, no human should be checking to see whether an items have a p-value less than .2 (one of the checks currently done by the Psychometrics department); a program containing all the error checks should be run on the data file and a list produced of all the items that have a difficulty of that value or less. From this point on, understanding the reason for the low p-value and determining

whether it is due to an error or not must be left to human judgment and the skills of the person investigating the issue. That is why it is important for a log to be kept of the issues explored and the findings; such a process will permit systematic review of the decisions made by the original investigator.

It is worth noting that we are not suggesting that a mechanical, systematic process be used entirely in place of human judgment. It always will be valuable to have a trained professional looking at the data to see whether it makes sense (indeed, we will suggest some additional such analyses in subsequent recommendations). However, there should be a set of checks that are done routinely on every data set, and these checks can and should be standardized and automated.

2. **Employ duplicate processing**. An effective way of reducing errors is to have another party process the same data that the original, responsible party is processing and determine when the final results are identical for both sets. This is especially important to do when judgment is an element in the final outcome, such as is true for scaling and equating. One possibility would be to create a second unit for Psychometrics within Harcourt; another would be to contract with an outside party. If the choice is to select an outside party, the decision about whom to hire could be made by Harcourt or Nevada. The panel and Harcourt staff discussed several ways this might be done, and generally agreed that an outside person would be best, since that person would have less knowledge of the procedures Harcourt used and therefore would have greater likelihood of using procedures that were different. There are three ways this could be done: (1) the outside party could use the same software and procedures as Harcourt, (2) use the same software, but not necessary the same procedures (for example, the outside party might make different decisions about which equating items to use and which to discard), or (3) use software completely independent of that used by Harcourt. In the first case, the outside person should get *exactly* the same results as Harcourt; in the second, highly *similar*, but not necessarily identical, results. With the third approach, there might be more differences between the two sets of results. There are advantages to each approach. The first would determine whether there were any mechanical errors in processing; the other two would highlight conceptual differences in approach that might be meaningful. The more independence the outside party has, the more confidence one would have in the *general* results if the two sets of findings were similar. However, there also is benefit to having the outside party duplicate Harcourt's results; that would provide more confidence in the accuracy of every student's and school's score. All three approaches have merit. The first likely should be done on every data set every year, while the other two likely should be done on occasion, with input from Nevada's Technical Advisory Committee, to determine the effect of decisions Harcourt has made about equating on the long-term stability of the processes. Harcourt and Nevada should discuss the merits of these various approaches and agree before the processing begins in any particular year on which should be used for that year's data.

3. **Check each year's results against other known results**. Until recently, Harcourt's checks about the reasonableness of data were relative to the data it was processing within that year. So, for example, they compared the number of students in the final reports to number of answer sheets that were originally processed. Another good check would be to compare the number of students included in the final reports this year to the number in the final reports last year. There are at least three ways that these checks could be done: comparing overall

results from this year to last year, comparing operational results to field test results, and having reports reviewed by local school personnel.

Harcourt already has begun to check the percentage of students scoring at each performance level this year compared to last year. They instituted this check after the most recent reporting error. Indeed, given that about one-third of the students were reported at a higher performance level than they actually earned, such a check certainly would have triggered further investigation before the results were released. Right now, the check is made at the state level only; the panel suggests that similar school-level checks be employed. The following SAS code suggests the kind of check that could be done simply:

```
PROC SORT DATA=NV.CRT500; BY SCHCODE;
PROC MEANS NOPRINT DATA=NV.CRT500;
  BY SCHCODE;
  VAR ELASS;
  OUTPUT OUT = SCHMEANS400
        MEAN = MEAN00
           N = N00;
PROC SORT DATA=NV.CRT501; BY SCHCODE;
PROC MEANS NOPRINT DATA=NV.CRT501;
  BY SCHCODE;
  VAR ELASS;
  OUTPUT OUT = SCHMEANS401
        MEAN = MEAN01
           N = N01;
DATA TEMP;
   MERGE SCHMEANS400 (IN = A) SCHMEANS401 (IN = B);
   BY SCHCODE;
   IF (A+B) = 2;
   IF MEAN00 > 0;
   IF MEAN01 > 1;
   IF 5 <= N01 < 20 THEN SIZE = 1;
   IF 20 <= N01 < 40 THEN SIZE = 2;
   IF 40 <= N01 < 80 THEN SIZE = 3;
   IF 80 <= N01 < 160 THEN SIZE = 4;
   IF 160 <= N01 THEN SIZE = 5;
PROC SORT; BY SIZE;
PROC CORR;
   VAR MEAN00 MEAN01;
PROC CORR;
   VAR MEAN00 MEAN01;
   BY SIZE;
PROC PLOT;
   PLOT MEAN00 * MEAN01;
   BY SIZE;
        RUN;
```

This program simply computes the mean for each school in two consecutive years, separates the schools by size (since larger schools tend to have more consistent performance from one year to the next), and then computes the correlations of results across years and prints a plot comparing the two years' results for each school. When this program was run on the CRT data from another state, we found these correlations:

Overall	.71

Schools with 5-19 students	.40
Schools with 20-39 students	.71
Schools with 40-79 students	.76
Schools with 80-15 students	.81
Schools with 160 students or more	.86

We also found that the mean was 3 scaled score points higher than the previous year and that the standard deviation was 2 scaled score points lower. All these results were consistent with the state's expectations, and therefore reinforced the belief that the results were accurate. The scatterplots showed that some schools with moderately high results had very low results the previous year; a follow-up showed that all these schools had either a very small number of students or a substantial reorganization from the previous year. All these checks contributed to increased confidence that the results were error-free.

Harcourt currently checks the difficulty of items in one year from those of the previous year for anchor items only. Another good check would be to compare the results of the remaining items this year with the results from the field testing done on those same items the previous year. Those items cannot be used for equating, since there are too many variables that can change from field testing to the inclusion of items in the operational forms, but there still would be information to glean from the comparison—and any significant discrepancies should be tracked down to determine the likely cause for the difference (in order to rule out error as the likely reason). Similarly, Harcourt should provide the p-values of the anchor items for both years to the Nevada Department of Education as soon as they are known. While a simple review of p-values will not definitely determine the overall amount of change in scores from one year to the next, they will indicate a reasonable range for such change. So, for example, if all the anchor items this year have p-values within 2 percentage points of their value last year, the Department should expect that any changes in student performance this year will be minimal compared to last year. That would serve both as a quality control check and an opportunity for the Department to prepare for the kinds of results it will be reporting to the public when the reports are released.

4. **Have local school people involved in the checking process.** The panel also believes that it would be productive to have a small number of schools review draft versions of their reports before final production. We would suggest that the Nevada Department of Education identify a limited number (perhaps 5 would be a reasonable number, especially if they were chosen to be diverse) of local school people who will honor the confidentiality of this assignment and the importance of completing it in a timely manner. These people should be convened to have the assignment explained and come to agreement on the checks they will do on the draft reports when received. On the date agreed to, Harcourt would post each school's reports as PDF files on their FTP site, the local school personnel would download their own reports through a password, conduct the checks they had agreed to, and report back to Harcourt about whether they observed problems that might indicate a system-wide problem. That is, the purpose of this check is to identify possible major errors, but not, for example, a coding error for an individual student. Harcourt would not be making changes for individual students at this point—that could not be done except at significant cost to the contract and with significant delays in the reporting schedule. This check would simply provide a final check before the decision was made to ship the reports that likely had already been produced.

5. **Have Nevada sign off on the test deck specifications.** Currently, Harcourt creates the specifications for the test deck using input from Nevada. The panel recommends that the process be more formal. Harcourt should propose Scoring Specifications and Research Specifications for Nevada's sign-off. Then, Harcourt should specify the parameters for the construction of the test deck from those specifications; Nevada should review and sign off on the plan for the test deck to ensure that a complete set of potential errors is checked.

6. **Apply the "test deck" concept to Psychometrics**. Currently, Harcourt constructs a test deck of answer sheets that is run through the entire front-end processing system of producing raw data files. The results from that processing are checked against the specifications that were used to create the test deck. That same concept could be applied to the Psychometric department. A set of specifications of possible outcomes could be created and those data processed before the "live" data are. Such a series of analyses would probe whether the system was working as intended.

7. **Make the Testmap even more central to the quality control process**. The panel was impressed with the concept of the Testmap and the procedures that Harcourt uses to ensure its accuracy. We also were impressed with some of the ways Harcourt is employing the Testmap to ensure quality, such as the automatic generation of file names. However, we believe the Testmap could be used more thoroughly throughout the process, from the generation of scoring specifications through the production of final reports. Automated production of many of these materials would significantly reduce the probability of error. For example, some field test items were inadvertently used as operational items in the scoring, scaling and equating of one test. If Testmap had been used, this error would not have occurred. Similarly, Harcourt uses the Testmap indirectly to create the scoring specifications; a "strand map" is derived from the Testmap and included in those specifications. But each time a process is used to convert the Testmap information into another format, there is potential for error. The panel recommends that these intermediate steps be eliminated, operating each process directly from the Testmap rather than from a derivative of the Testmap file.

8. **Employ a systematic close out procedure at the end of each psychometric processing cycle.** Multiple data and program files are created as the Psychometrics department analyzes current year results. Many of the files are temporary outputs from various steps in the process. At the end of the cycle only the final scaling, equating, and score tables are relevant for processing the current year's results. Harcourt has taken steps to standardize the naming conventions for these files to help ensure that the appropriate files used. We recommend that at the end of the processing cycle that the final files are moved to a secure archive with appropriate quality assurance checks. All temporary files should be deleted to further ensure that only appropriate files are accessible thereby reducing the likelihood that a temporary file is mistakenly used in the process. It is important to note that there are two specific sets of data that might be affected. The first includes the scoring, scaling and equating files; the second includes the "final" set of results that Harcourt releases to Nevada. Harcourt already is moving to implement a formal closeout of the first set; that process needs to be completed. For the second set, there should be a mutual sign-off, but since Nevada may change those "final" data once they have received them (based on, for example, appeals that districts might make to the state), Harcourt should not be responsible for archiving those files.