Digital Test Decks® and What They Can Do for Your Forms Data Capture System

A White Paper

K. Bradley Paxton, Ph.D.
ADI, LLC
March 28, 2007

Introduction

A major problem with capturing data from forms filled out by respondents is measuring the accuracy and efficiency of the system. This is true whether traditional "heads-down" keying from paper (KFP), "heads-up" keying from image (KFI), and/or handprint Optical Character Recognition (OCR/ICR) is capturing the data.

It is a fundamental fact that in order to improve the accuracy and efficiency of a system, it is necessary to be able to measure these performance factors. Using our new, patent pending Digital Test Deck® technology, you can now assess your forms processing system more easily and accurately than ever before.

Forms processing technology can be a bit complicated, so in this white paper, we attempt to provide a simple explanation of the basics, and how Digital Test Deck® technology can be used to save you a lot of money while insuring system accuracy.

What is Forms Data Capture?

The term "Forms Data Capture" can mean a lot of things, including just management of business documents, but here we take it to mean the capturing of handprinted data from pre-printed paper questionnaires and putting the data into a computer file. An example with which we are quite familiar (because we worked on it for eight years) is the Year 2000 Decennial Census, where automatic recognition was used to process 120 million forms in 100 days.

Organizations doing forms data capture usually cycle through the following three phases as they improve their data capture accuracy and efficiency:

- Traditional "heads-down" keying from paper (KFP)
- "Heads-up" keying from image (KFI)
- Handprint Optical Character Recognition (OCR/ICR)

We will discuss each of these three phases in order.

Traditional "heads-down" keying from paper (KFP)
This time-honored approach involves human keyers sitting in front of a computer terminal, and looking down at a form placed on a rack. They read the data placed on the form by the respondent, and manually key this data into the computer using a KFP software package. People can be amazingly good at this, but they can also be amazingly bad. This is the essential problem with this approach: Are they good or bad? People are not very consistent when performing routine tasks for long periods of time. A major source of error with KFP is placing the data in the wrong field.

In order to help explain the problem, I have often used a teaching aid originally developed by Professor Rickmers at Rochester Institute of Technology, and shown below:

Count the number of "F"s in the following sentence:

FINISHED FILES ARE RESULT
OF YEARS OF SCIENTIFIC
STUDY COMBINED WITH
THE EXPERIENCE OF YEARS.

When I use this example in a talk and ask for a show of hands, I get answers ranging from two to seven (the correct answer is six). The point of this example is merely that we, as humans, are not really very good at observing the sorts of details that computers do extremely well. We do not mean to disparage the very excellent keyers who do exist out there, either: only to indicate that since humans are involved, errors will occur in unpredictable ways. Of course, keyers are expensive, and turnover, consistency, and training are major issues.

"Heads-up" keying from image (KFI)
In this approach, an electronic scanner first scans the forms, and the electronic image of the form sent to a computer screen, along with fields wherein the data is to be keyed by the human keyer. The keyer is looking straight ahead at the screen at all times, and hence the name "heads-up" keying. The software presents the fields to be keyed in an orderly fashion, and tends to greatly reduce the incorrect field placement problem mentioned above for KFP. Key from image is faster than KFP, and more accurate, however, it still involves humans, who are costly and make mistakes.

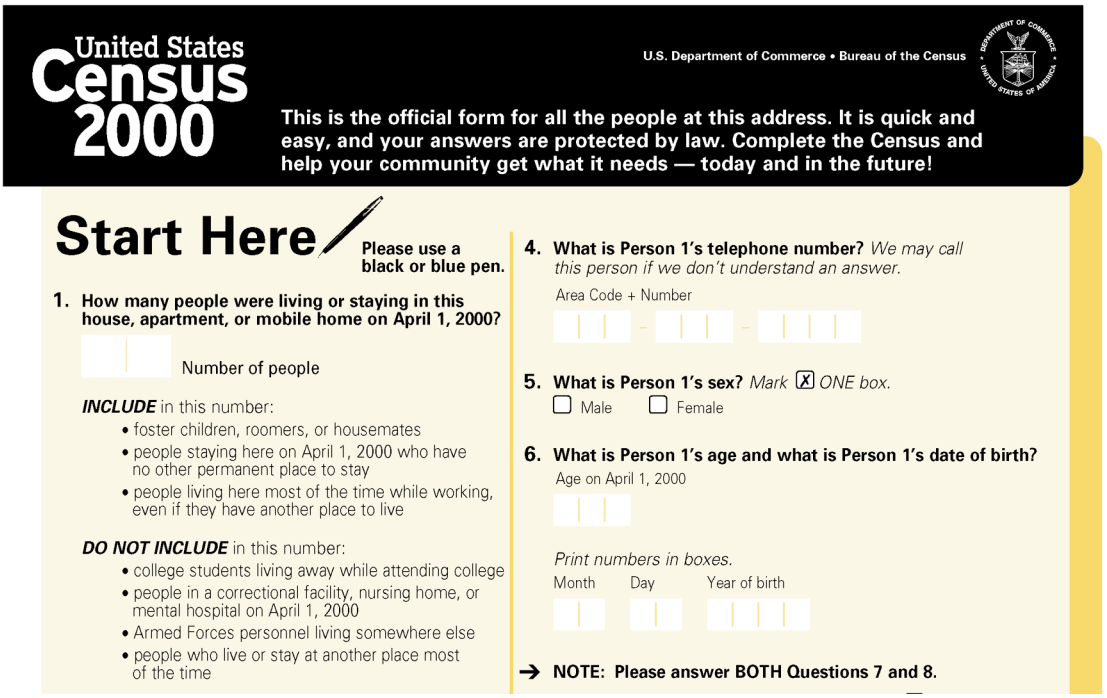
Handprint Optical Character Recognition (OCR)
This approach, sometimes called Intelligent Character Recognition (ICR), employs sophisticated software, often involving neural network technology, to automatically capture the handprint data from the electronic image of the form. These systems do the majority of the work in a given data capture task, but cannot do it all due to the inherent sloppiness and variability of human handprint. Therefore, some write-in fields, which are difficult to recognize with high confidence, are said to be "rejected" by the OCR engine,

and sent to KFI for data capture. The accuracy of modern OCR systems (for the "accepted" fields) is very high, and considerably better (and faster) than human keying.

What is a Digital Test Deck® Anyway?

In simplest terms, a Digital Test Deck® is a stack of forms containing simulated human handprint, which look just like real forms even though they were printed by a digital press and contain perfectly known source data. Using these decks, one can test their forms data capture system for accuracy and efficiency, regardless of the technology used to do the data capture. More specifically, one can assess the state of their system now, and using the Digital Test Deck® technology, gain confidence in progressing through the phases mentioned above, resulting in lower costs and greater verified accuracy.

As an example, here is a portion of the Year 2000 Decennial Census "short" form:



United States Census 2000

U.S. Department of Commerce • Bureau of the Census

This is the official form for all the people at this address. It is quick and easy, and your answers are protected by law. Complete the Census and help your community get what it needs — today and in the future!

Start Here — Please use a black or blue pen.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2000?

Number of people

INCLUDE in this number:
- foster children, roomers, or housemates
- people staying here on April 1, 2000 who have no other permanent place to stay
- people living here most of the time while working, even if they have another place to live

DO NOT INCLUDE in this number:
- college students living away while attending college
- people in a correctional facility, nursing home, or mental hospital on April 1, 2000
- Armed Forces personnel living somewhere else
- people who live or stay at another place most of the time

4. What is Person 1's telephone number? *We may call this person if we don't understand an answer.*

Area Code + Number

5. What is Person 1's sex? *Mark* ☒ *ONE box.*

☐ Male ☐ Female

6. What is Person 1's age and what is Person 1's date of birth?

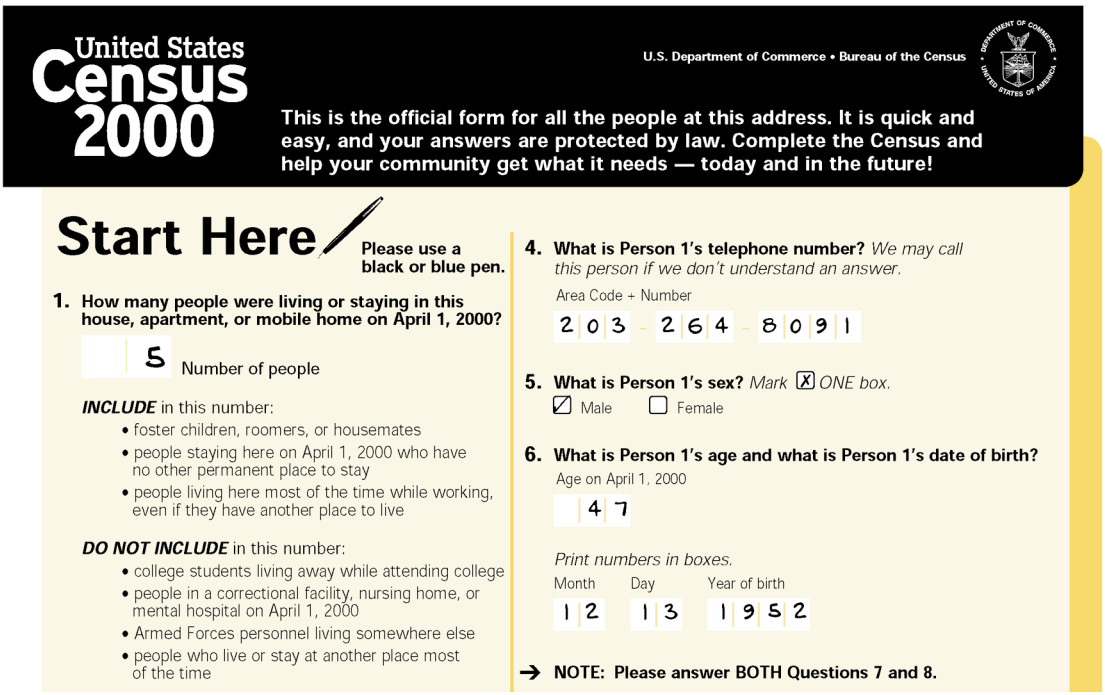
Age on April 1, 2000

Print numbers in boxes.

Month Day Year of birth

→ NOTE: Please answer BOTH Questions 7 and 8.

There are places on this (blank) form, called fields, where the respondent is asked to print the answers to questions posed by the Census Bureau. When the respondent completes these fields, the form might look like this:



United States Census 2000

U.S. Department of Commerce • Bureau of the Census

This is the official form for all the people at this address. It is quick and easy, and your answers are protected by law. Complete the Census and help your community get what it needs — today and in the future!

Start Here Please use a black or blue pen.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2000?

5 Number of people

INCLUDE in this number:
- foster children, roomers, or housemates
- people staying here on April 1, 2000 who have no other permanent place to stay
- people living here most of the time while working, even if they have another place to live

DO NOT INCLUDE in this number:
- college students living away while attending college
- people in a correctional facility, nursing home, or mental hospital on April 1, 2000
- Armed Forces personnel living somewhere else
- people who live or stay at another place most of the time

4. What is Person 1's telephone number? *We may call this person if we don't understand an answer.*

Area Code + Number

2 0 3 - 2 6 4 - 8 0 9 1

5. What is Person 1's sex? *Mark* ☒ *ONE box.*

☑ Male ☐ Female

6. What is Person 1's age and what is Person 1's date of birth?

Age on April 1, 2000

4 7

Print numbers in boxes.

Month 1 2 Day 1 3 Year of birth 1 9 5 2

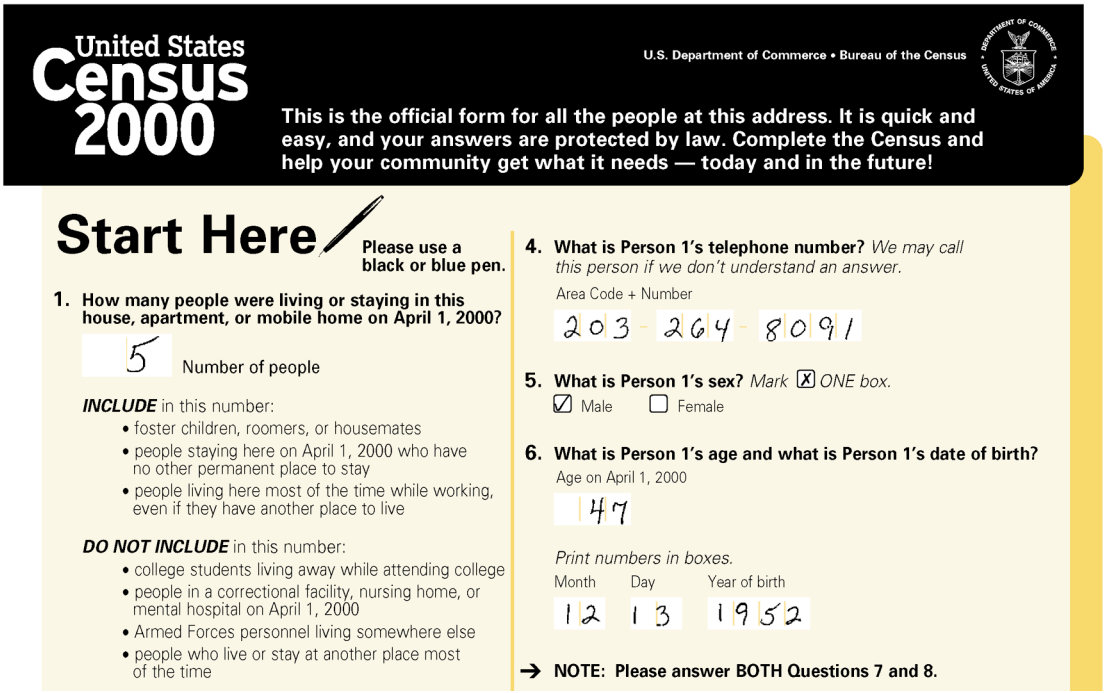
→ NOTE: Please answer BOTH Questions 7 and 8.

Most people would say this was an actual Census form image, (albeit a rather neat one), but it was actually created using a handprint font on a computer. This is an early example of what we mean by a digital test form. A suitable number of different digital test forms would constitute a Digital Test Deck®.

The basic properties of a Digital Test Deck® as defined above are:

- Looks and feels like a real form with handprinted responses, but really printed on a high quality Digital Color (or black & white) Press.
- Form content designed to test critical System aspects.
- Reproducible as required.
- Compliments, but does not replace forms with "real data" content.
- Consistent test input.
- Test the data capture system "end-to-end"
- Know the "truth" of the data fields perfectly!

The fourth bullet indicates the fact that, using handprint fonts, one obtains a rather excessively "neat" simulated form, and so the form is of limited use in actually measuring OCR data capture quality. However, using actual handprint "snippets" in the creation of the Digital Test Deck® gives a very realistic appearance, as shown below:



United States Census 2000

U.S. Department of Commerce • Bureau of the Census

This is the official form for all the people at this address. It is quick and easy, and your answers are protected by law. Complete the Census and help your community get what it needs — today and in the future!

Start Here Please use a black or blue pen.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2000?

5 Number of people

INCLUDE in this number:
- foster children, roomers, or housemates
- people staying here on April 1, 2000 who have no other permanent place to stay
- people living here most of the time while working, even if they have another place to live

DO NOT INCLUDE in this number:
- college students living away while attending college
- people in a correctional facility, nursing home, or mental hospital on April 1, 2000
- Armed Forces personnel living somewhere else
- people who live or stay at another place most of the time

4. What is Person 1's telephone number? *We may call this person if we don't understand an answer.*

Area Code + Number

203 - 264 - 8091

5. What is Person 1's sex? *Mark* ☒ *ONE box.*

☑ Male ☐ Female

6. What is Person 1's age and what is Person 1's date of birth?

Age on April 1, 2000

47

Print numbers in boxes.

Month 12 Day 13 Year of birth 1952

→ NOTE: Please answer BOTH Questions 7 and 8.

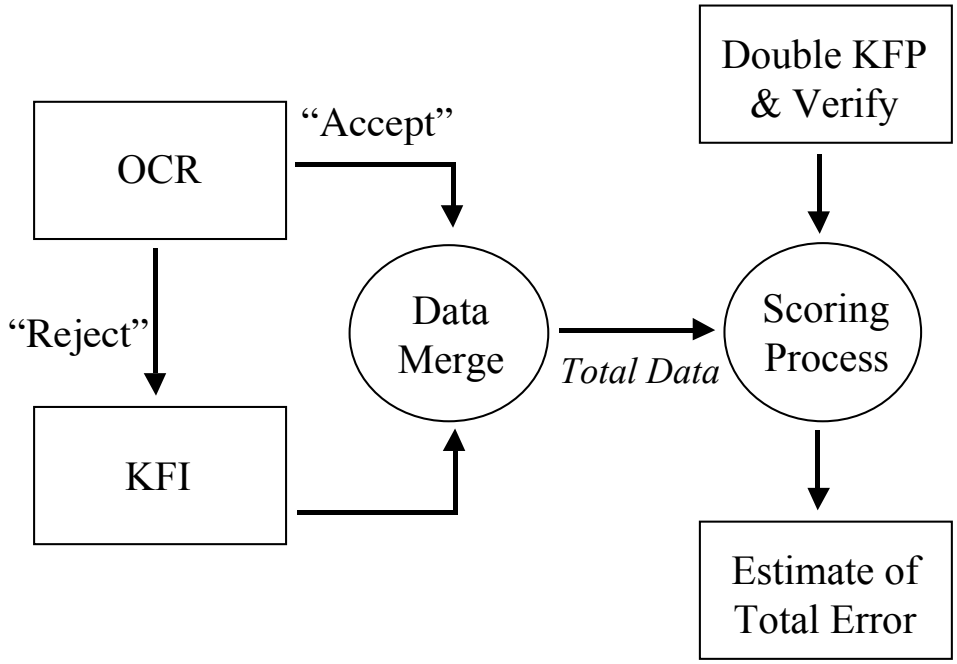
By "snippet" we mean an image clip containing a handprint character or field. Using forms with actual handprint snippets, as shown above, one can actually test forms processing OCR data quality with assurance of realistic results. We draw upon a massive collection of actual human handprint snippets to form our test decks, and in addition, can make them look like they were written with various marking instruments, such as pencil, black rolling ball, blue ballpoint, etc.

What are the uses of a Digital Test Deck®?

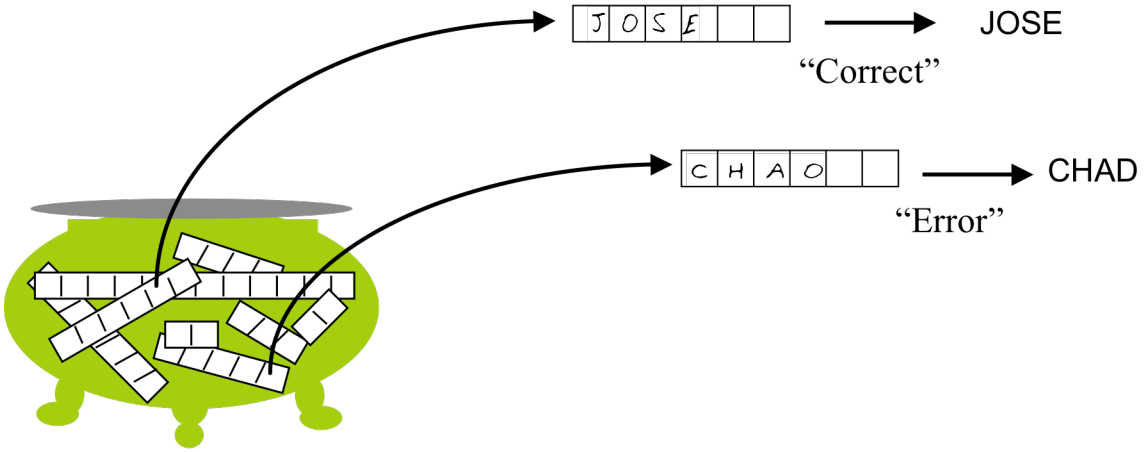
Generally, we find the following uses for a Digital Test Deck®:

- Verify correct operation of critical System Requirements
- Establish measurable System performance baseline.
- Test System operation at each software/hardware change.
- Test consistency of System between scanners, sites, over time.
- Obtain a statistically adequate volume of test materials
- Test daily operational readiness before scanning.

How to Measure Accept Rate and Accuracy of a Forms Processing System
In the diagram below, we show a typical forms processing system which uses automatic recognition (OCR) to do the bulk of the data capture workload, and KFI for data capture of those rejected fields for which the OCR system is not confident. Also shown are the steps one might take to measure the accept rate and accuracy of the system. In this case, we indicate that an expensive and laborious process is done on the test forms wherein the forms are keyed twice from paper, and discrepancies verified and corrected by a third person.



The testing proposition can be thought of as in the following picture:



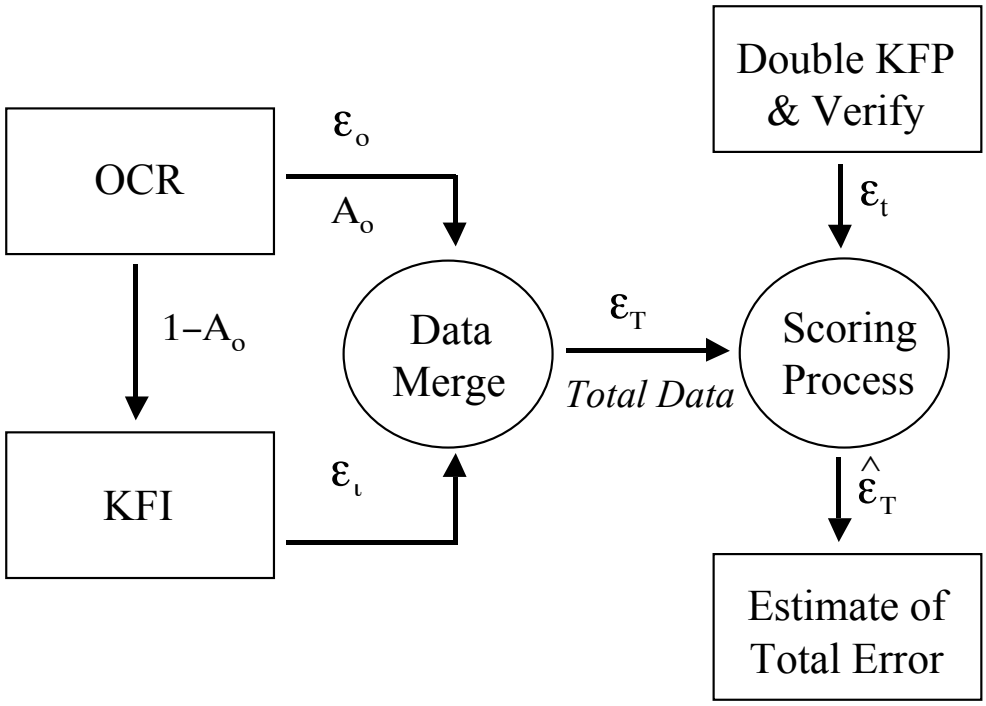
Universe of Fields

You think of all the fields in all the forms being tested as being in a large vat, and some are pulled out one at a time for testing. If the handprint was J O S E and the resultant ASCII was JOSE, then that would be a correct field, under what we call a "hard match", meaning each and every character is correct in a field.

On the other hand, if the handprint was C H A O and if the resultant ASCII was CHAD, that would be an error using the hard match criterion. Other criteria for correctness are possible which are beyond the scope of this paper.

Another point to realize here is that we are talking about accuracy results "at the field level", which is what one usually prefers to use as it relates most readily to one's business needs. The use of character level accuracy is much less satisfactory, although it gives much higher accuracy numbers; hence it's popularity.

If we define some mathematical terms in this process, we get the revised diagram shown below:



Here, the total error estimate is:

$$\hat{\varepsilon}_T = \varepsilon_o A_o + \varepsilon_\iota (1-A_o) + \varepsilon_t$$

ε_o = OCR Error

A_o = OCR Accept Rate

ε_ι = KFI Error

ε_t = "truth" Error

$\hat{\varepsilon}_T$ = Estimate of Total Data Error

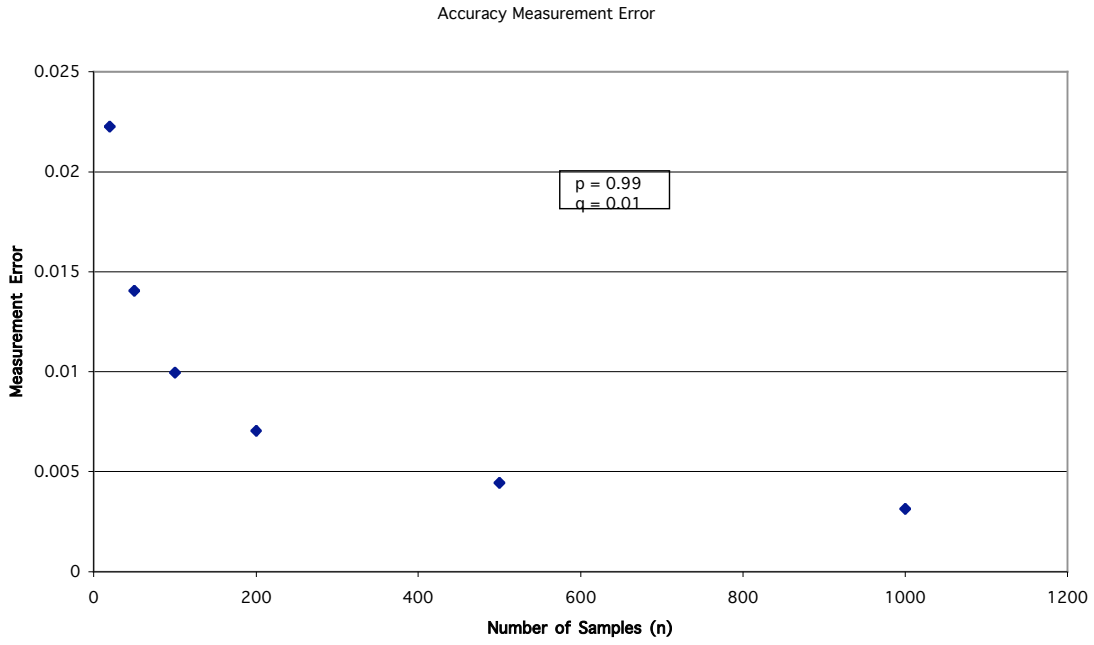
This model is very useful in analyzing data capture systems, and can, for example, allow the estimation of total system error from measurements of OCR and KFI sub-systems. Of course, when one uses a Digital Test Deck®, the "truth" error is zero, not to mention you save a ton of work since you don't have to create a large test deck manually.

How to Associate Measurement Error with Number of Samples

If we are attempting to measure an accuracy in the neighborhood of p = 99%, say, corresponding to an error rate of q = 1%, then we find the following relationship to describe the one-sigma error in the estimate as:

$$\sigma = \sqrt{\frac{pq}{n}},$$

where n is the number of samples. Plotted up it looks like:



Using this simple method, one can decide how many samples one needs to use to obtain the desired level of quality in estimating your system accuracy.

One of the surprising aspects of sampling is that the required number of samples is on an absolute basis, not a percent of the population, per se. A more refined formula may be used if the population size is small relative to the sample size.

Another fundamental, and sometimes troubling fact is that the required number of samples rises very rapidly as the error we are trying to measure gets smaller. This means, of course, that we have to work harder the better we get!

Fortunately, a large volume of realistic test data is one of the benefits of using a Digital Test Deck®.

Test Materials for Forms Processing Systems
We have had extensive experience with the following six types of test materials that may be used to test forms processing systems:

- Blank Forms
- Forms hand-marked by volunteers
- Real forms filled out by respondents
- Images of real forms on CD-ROM
- Lithographically printed forms with simulated respondent marks
- Digitally printed forms with simulated respondent marks

Each of these types of test materials has a purpose, and has advantages and disadvantages. By a suitable combination of these materials, one can devise tests to cover all testing needs.

Blank Forms

Advantages:	Inexpensive.
	Exercise scanners, sorters, paper handling.
Disadvantages:	Don't exercise recognition subsystems or keying.

Forms hand-marked by volunteers

Advantages:	Content under Project control.
	Real handprint.
Disadvantages:	Tend to get cute, pathological, or non-realistic inputs.
	Only one test set available – not reproducible.
	Human error makes them quite variable.
	Critical to carefully instruct volunteers.
	Difficult to obtain accurate "truth".

Real forms filled out by respondents

Advantages:	The only "real" data.
	Used for measuring OCR/KFI Data Quality.
Disadvantages:	Only one precious "golden" deck – keep locked up.
	Tends to get beat up and change over time.
	Obtaining "truth" is arduous and never "done".
	Some fields are, at best, ambiguous.
	Truth error plus ambiguity makes a "floor" below which the Data Capture System cannot go, even if it's perfect.

Images of real forms on CD-ROM

Advantages:	Consistent source of images for processing tests.
	Never wear out.
	Reproducible and inexpensive.
Disadvantages:	Don't test scanner-input subsystem, image processing or paper handling.
	Don't fit nominal workflow.

Lithographically printed forms with simulated respondent marks

Advantages:	Inexpensive.
	Reproducible.
	Useful to test most of system, and some OCR/KFI.
Disadvantages:	Limited data variation.
	Not useful for serious OCR/KFI quality assessment.

Digitally printed forms with simulated respondent marks

Advantages:	Complete control over content.
	Reproducible – consistent.
	Moderate to high data variability.
	Can test everything *but* OCR/KFI quality if handprint fonts are used.
	Can test OCR/KFI data quality if real handprint snippets are used.
	Know "truth" perfectly, for once.
Disadvantages:	More costly than lithographic printing.
	Handprint fonts are "too neat"?
	Handprint snippets not "real" marks (but can be very close),

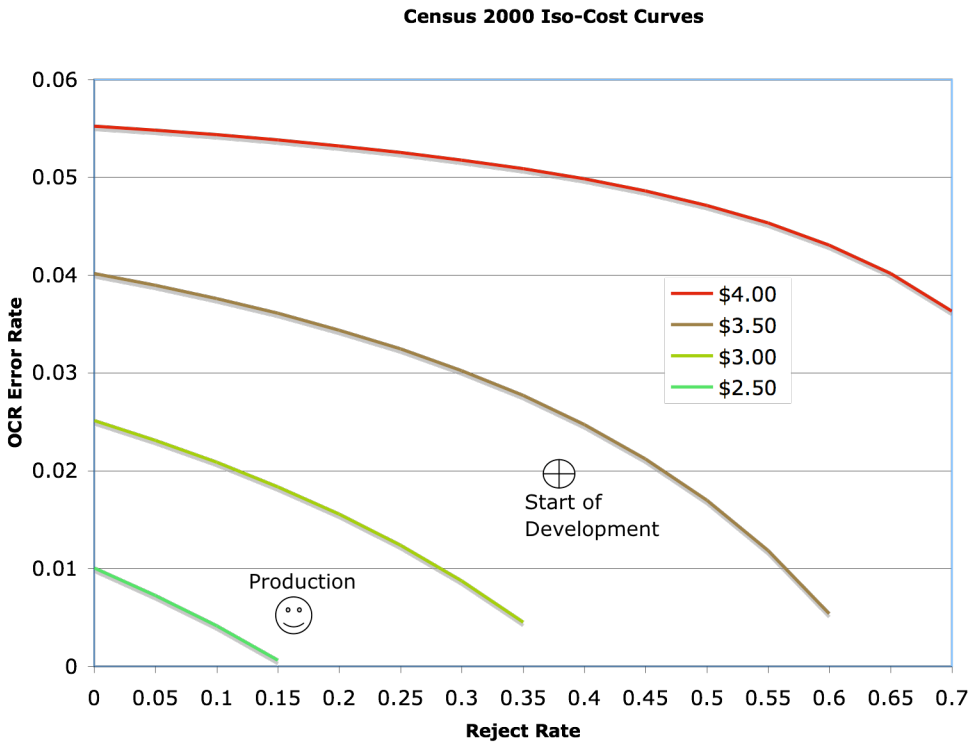
A Digital Test Deck® made with realistic actual handprint image snippets is the "ultimate" in test materials, and allows you to test your entire forms processing system.

We call the kind of testing one can do with a Digital Test Deck® "outside-in" testing. The basic idea is that if a perfectly known input is inserted into the system, and (mostly) the correct answers come out, then it is unlikely that there is anything seriously wrong in between. An alternative is "inside-out" testing, wherein one may measure, say, the lamp intensity on the left-hand side of the scanner, or the speed variation of the transport, etc. The problem with this approach is that one may literally fill up file cabinets with data in this manner, and it will be the thing that is <u>not</u> tested that causes your system to fail. Our approach to "outside-in" testing is cost-effective, accurate, and consistent.

A Simple Cost Model Example for a Forms Processing System

We have created several cost models applicable to making decisions about forms processing data capture systems that vary in complexity. Many factors come into play when estimating the costs of data capture. Some of these are: the amortized cost of equipment and software, labor, facilities, etc. Some factors are technical in nature, such as: OCR Error vs. Reject Rate, forms volume and complexity, keying speeds and costs, keying accuracy, and finally, the cost of an error in data capture downstream in the business process.

For our purposes in this white paper, we show a typical cost model result, based on parameters and test data from the U.S. 2000 Decennial Census.



In the above graph we plot OCR/ICR error rate on the vertical axis, and OCR/ICR reject rate on the horizontal axis. The curved lines are lines of constant system cost to process one, somewhat complex, form. The way a plot like this is used is to first run an OCR/ICR data capture test to establish a baseline performance; here we show a crossed circle at 0.02 (2%) error rate and 0.37 (37%) reject rate. This represents the Census 2000 data capture system at the start of the development work, and one can read off a total cost per form of less than $3.50, but greater than $3.00; about $3.30. Of course, used in the detailed model are many other parameters, such as keyer costs, etc. mentioned above.

After a series of detailed tests and process improvements, the 2000 Census system was put into production, and the results measured at that time are shown by the "smiley face", at about 0.006 (0.6%) error rate, 0.16 (16%) reject rate, and near a total cost per form of about $2.60.

The difference between these two points in terms of cost is about $0.70 per form. Since the Decennial Census 2000 processed over 100 million forms, this corresponds to a savings of about $70,000,000.

This example shows the basic idea of how good measurements of data system performance can be related to total cost per form. Depending on the particular technical and business parameters of a given data capture system, the iso-cost curves will move around somewhat, but the fundamental idea remains the same. More detailed results may also be obtained, depending on one's needs.

For more information, contact:

Dr. K. Bradley Paxton
ADI, LLC
200 Canal View Blvd.
Rochester, NY 14623
(585) 239-6057
brad.paxton@adillc.net
www.adillc.net