Human-Based vs. Optical Scanning Methods of Data Entry

A Comparison of Data Quality, Cost, and Efficiency

International Field Director's Conference – May 2009

Lisa Klein and Christopher Huard

UWSC

University of Wisconsin Survey Center

www.uwsc.wisc.edu

Outline

- Background
  - CASES Entry and Delivery
  - Optical Scanning Technology
- Experiment Goals
- Results
  - Data Quality
  - Cost Analysis
- Recommendations

Background and Research Questions

- The UWSC typically enters data via CASES
- In 2008, the UWSC inherited optical scanning technology (Teleforms) from another UW department
- Research questions:
  - Can optical scanning technology process data in a cost-effective manner ***without compromising data quality?***
  - How does the total cost differ between methods?
  - What is our best practice recommendation about choosing an entry mode?

Experimental Design

- Parallel data deliveries conducted on the Survey of Washington Physicians
  - 4 page mail questionnaire of doctors in Washington
  - Primarily close-ended response fields
  - Single-entry of returned questionnaires
  - Survey not initially designed to be Teleforms-compatible, but by coincidence was able to be processed using Teleforms
  - Same 150 cases manually entered and delivered via CASES, then optically scanned and delivered via Teleforms
  - Data from both deliveries output in SPSS

CASES Data Entry and Delivery Fundamentals

- The vast majority of the UWSC’s returned mail questionnaires are processed via human-based entry
  - Interviewer manually enters responses from each questionnaire into a programmed CASES instrument
  - Cases can be single- or double-entered (depending on budgetary constraints)
  - Interviewers can leave notes about unclear answers or respondent marginal comments at any item, or at the end of the instrument

Human-Based Entry and Delivery Fundamentals

- This process of data entry yields high quality data and accurate entry.
- However, human-based data entry has a variety of associated costs, including:
  - Staff training
  - Instrument programming
  - Associated licensing fees
  - Quality control and data checking
  - Data delivery (programmer and project director time)

Optical Scanning Technology (Teleforms)

- Teleforms is designed to process a high volume of surveys through scanning technology
- OCR (optical character recognition) software, which is a technology that uses artificial intelligence to translate images of writing into machine-editable text
- Capable of creating data sets in a number of different formats including Excel, Access, SPSS, etc.

Teleforms Scanner

Preparing Cases for Scanning

- Initial visual scan by human operator
  - Pencil
  - Skip-pattern problems
  - Torn or ripped surveys
- Organizing surveys in correct orientation
- Dividing surveys into batches

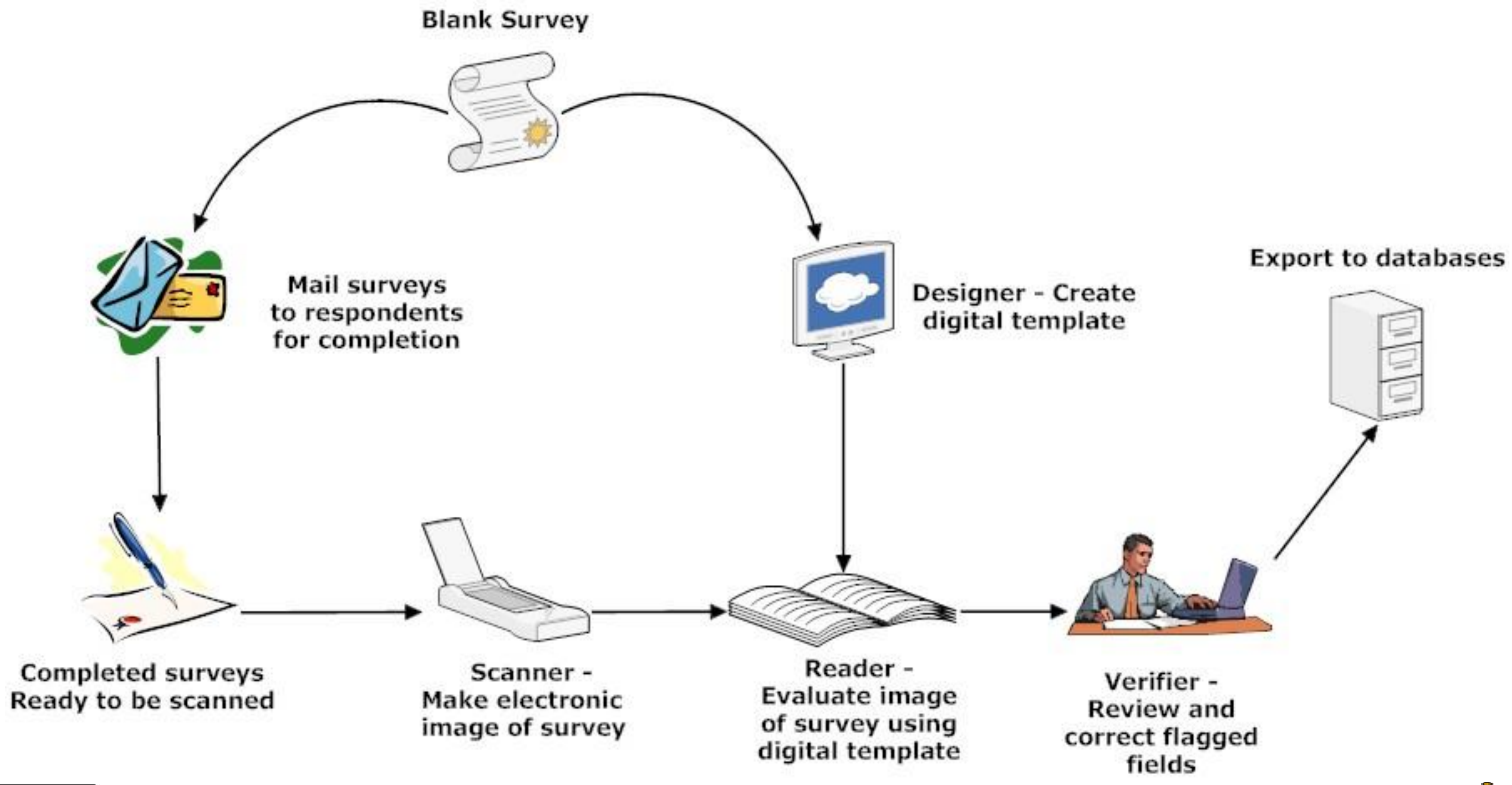
Phases of Teleforms

- **Designer** – create digital template using tools to overlay response fields
  - Define field values and database export
- **Scan Station** – completed surveys are scanned which create electronic images
- **Reader** – evaluates completed surveys against digital template
  - Works behind the scenes

Phases of Teleforms (cont.)

- **Verifier** – in this last phase, a human operator is presented with the electronic image of a survey
  - Verifier is the critical quality control function of Teleforms
  - The human operator makes decisions regarding fields that did not meet the specifications defined in Designer
  - For example, a field would be flagged for review if the respondent was only supposed to select one answer choice but for some reason they selected two
  - After the operator manually reviews each flagged field the data are ready to be exported to a database

Teleforms Process: Start to Finish


Blank Survey
Mail surveys to respondents for completion
Designer - Create digital template
Export to databases
Completed surveys Ready to be scanned
Scanner - Make electronic image of survey
Reader - Evaluate image of survey using digital template
Verifier - Review and correct flagged fields

Evaluation Criteria

- Is there a difference in data quality?
  - Missing data
  - Miscoded data
  - Accuracy of coding open-ended response items
  - Recognition of respondent marginal comments
- What are the associated costs of each method?
  - Staff hours/wages
  - Hardware costs/licensing fees
  - Training time
  - Programming/set-up
  - Data review/corrections
  - Quality Control

Close-Ended Responses: Missing Data and Miscoded Data

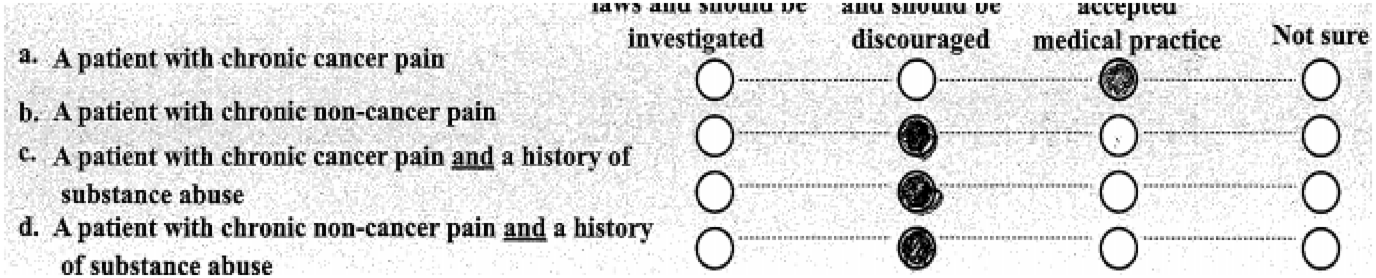
	Teleforms Processing Error	Human-Based Entry Error
Missing Data	33	0
Miscoded Data	7	20
Total Errors	40	20
Error Rate	.35%	.17%

Close-Ended Responses: Missing Data and Miscoded Data

	Teleforms Processing Error	Human-Based Entry Error	Pencil-Caused Error
Missing Data	4	0	29
Miscoded Data	7	20	0
Total Errors	11	20	29
Error Rate	.09%	.17%	.25%

Missing Data – Pencil Errors

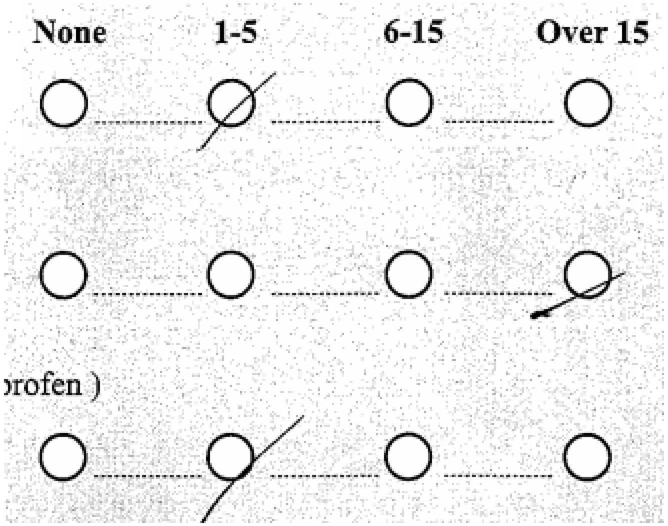
- The scanner was unable to detect most pencil marks and was therefore coded as blank (though testing suggests that with certain configurations, pencil recognition can be improved)
- Even though the circles are completely filled in the scanner has a hard time recognizing them



	laws and should be investigated	and should be discouraged	accepted medical practice	Not sure
a. A patient with chronic cancer pain	○	○	●	○
b. A patient with chronic non-cancer pain	○	●	○	○
c. A patient with chronic cancer pain <u>and</u> a history of substance abuse	○	●	○	○
d. A patient with chronic non-cancer pain <u>and</u> a history of substance abuse	○	●	○	○

Missing Data – Field Not Filled in Enough

- Sometimes the Respondent’s markings are too vague for Teleforms to detect even when using pen

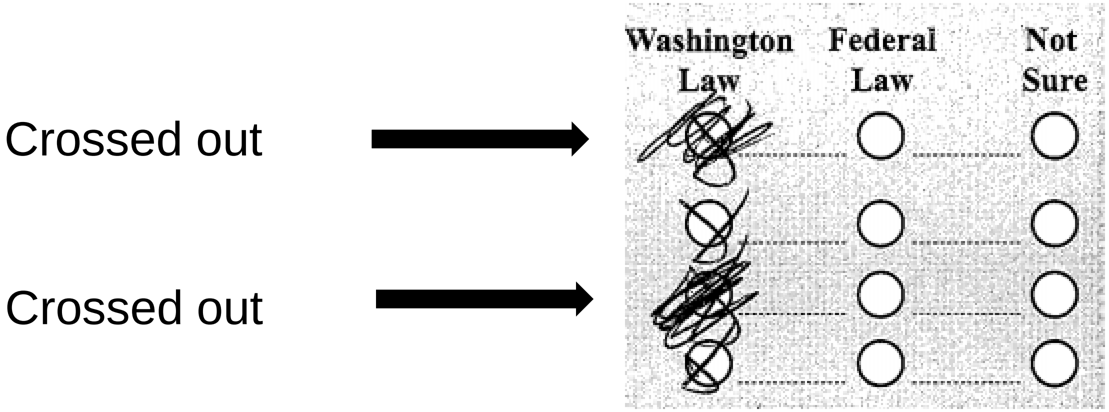


None	1-5	6-15	Over 15
○	[✓ ○]	○	○
○	○	○	[✓ ○]
rofen)			
○	[✓ ○]	○	○

- When Respondents mark too lightly Teleforms incorrectly counts these fields as blanks

Miscoded Data - Bubble Filled and X’d Out

- Teleforms cannot tell when a Respondent has filled in a bubble but then crosses it out

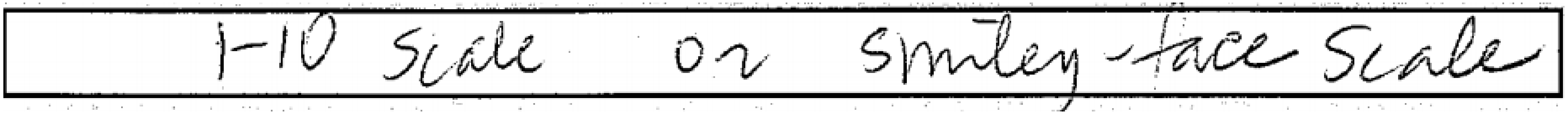


- Teleforms mistakenly counts these items as being filled in, even though it is clear they have been crossed out
- It is only looking for *any* marking within a field

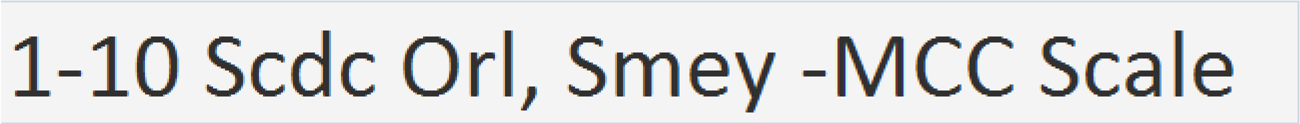
Open-Ended Response Errors

- There were 61 occurrences in which the Respondent answered an open-ended field
    - **Without review** Teleforms only correctly interpreted the Respondent’s answer 2 times
    - However, **with review**, there were virtually no errors
- Teleforms did a better job of interpreting numerical responses (55.3% accuracy), but is still far too error-prone to utilize without review
- Because of this huge discrepancy we make it a rule to set up open-ended fields in such a manner that they are always reviewed

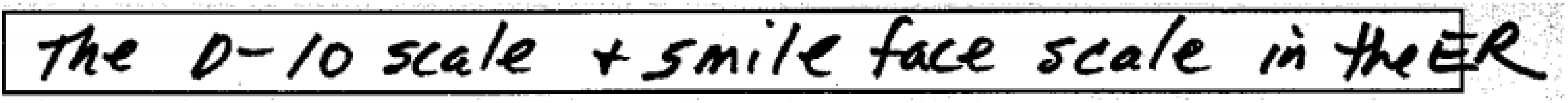
Teleforms – Open-Ended : Examples



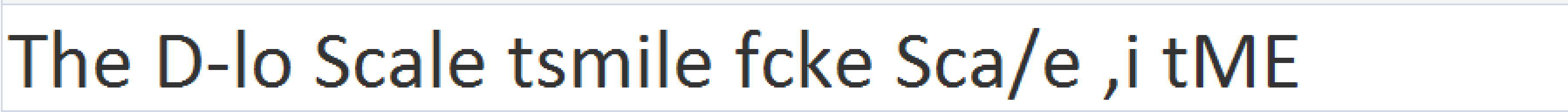
1-10 Scale or smiley-face Scale



1-10 Scdc Orl, Smey -MCC Scale



The 0-10 scale + smile face scale in the ER



The D-lo Scale tsmile fcke Sca/e ,i tME

Marginal Comments

- Since there is no way to tell when and where a Respondent will make a marginal comment, there is no way to program Teleforms to capture this data
- To capture all marginal comments a data entry operator would have to manually look over each survey, and enter the responses
- This is one weakness of Teleforms

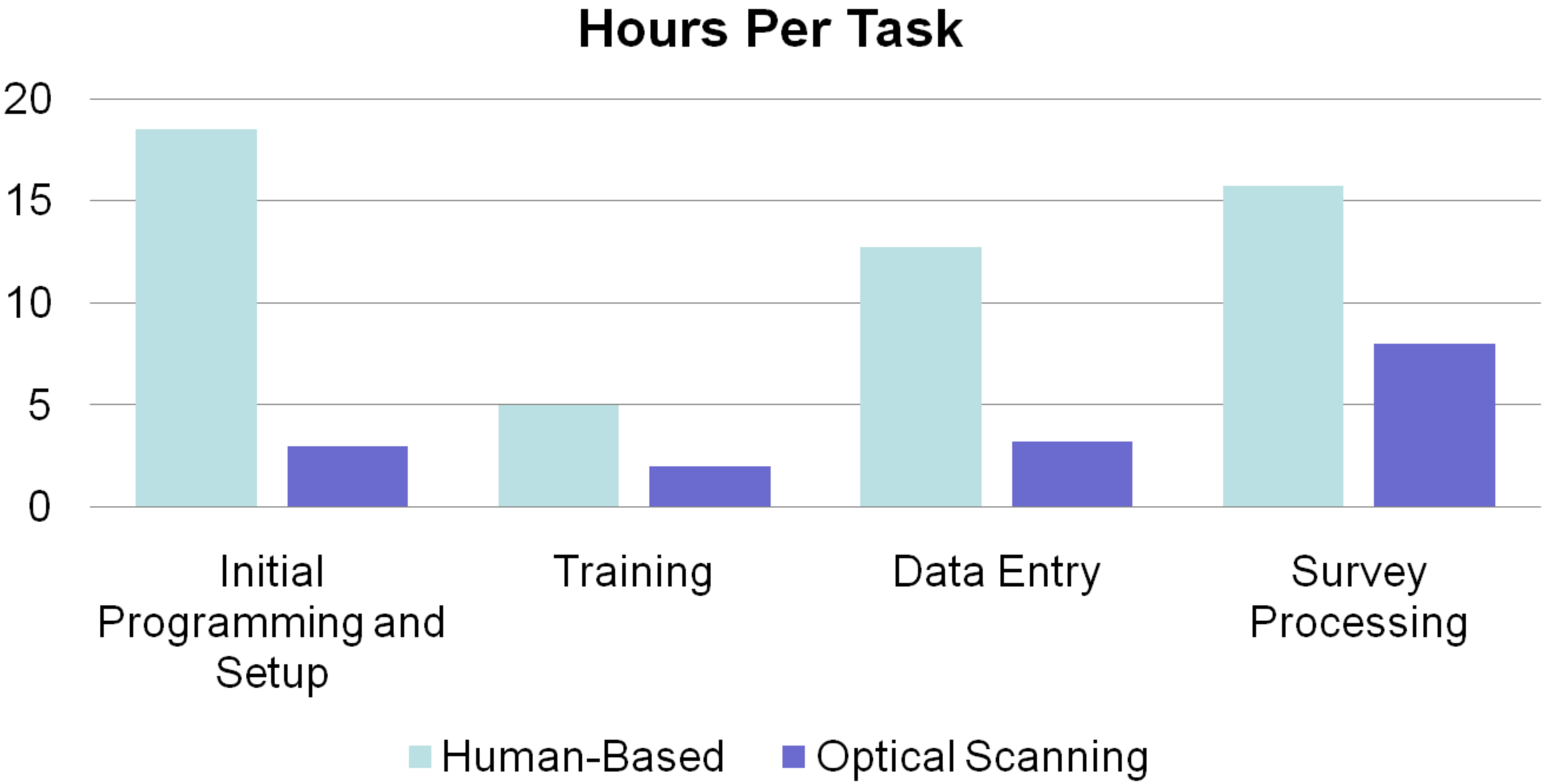
Efficiency Comparison: Programming and Development

	Human-Based	**Optical Scanning**
Programming Time & Hardware Setup	10.5 programmer hours 11 debugging &review hours	3 hours
Staff Training and Protocol Development	5 hours	2 hours
Total Programming and Development Time	**26.5 hours**	**5 hours**

Efficiency Comparison: Survey Processing

	Human-Based	**Optical Scanning**
Staff Data Entry Time	12.75 hours	3.25 hours
Processing Time (clean-up, quality control)	3 programmer hours 12.75 supervisory hours	8 hours
Total Processing Time	**28.5 hours**	**11.25 hours**

Relative Efficiency



Relative Cost Per Survey

150 cases

	Human-Based	Optical Scanning
Programming, Debugging, and Hardware Setup	$4.50	$.43
Staff Training	$1.01	$.41
Staff Data Entry	$1.03	$.47
Staff Data Processing	$3.27	$1.14
Total Per Survey Cost	**$9.81**	**$2.45**

Scaled to 1500 cases

Total Per Survey Cost	**$2.15**	**$.76**

Cost Considerations

- Per-survey estimates do not include costs fixed across modes (such as instrument review and printing)
- Start-up costs for TeleForms were significant (~$10,000 in staff time and tech consultations), though per-year costs of TeleForms and CASES are roughly equal

Our Recommendation

- Scannable survey entry is a cost-effective option when the survey:
  - Is relatively short in length
  - Has few or no open-ended text items
  - Was designed with machine requirements in mind (ours was not)
    - Pencil
    - Skip-pattern complexity
    - Item definitions

Our Recommendation

- Scannable survey entry may not be a viable option if:
  - Staff resources aren’t available to integrate new technology
  - The survey has many open-text fields requiring review
  - The survey has complex, potentially error-prone skip-patterns
  - Design elements of the survey are not scanning-compatible

Questions?

Lisa Klein, Project Director: lklein@ssc.wisc.edu

Christopher Huard, Project Assistant: chuard@ssc.wisc.edu