2011

Balancing truth error and manual processing in the PDQ system

Douglass Huang

Balancing Truth Error and Manual Processing in the PDQ System

by

Douglass Huang

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science
in Computer Science

Supervised by

Professor Roger S. Gaborski, Ph.D.
Department of Computer Science
B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York
August 25, 2011

Approved by:

Roger S. Gaborski, Ph.D., Professor
Thesis Advisor, Department of Computer Science

Peter G. Anderson, Ph.D., Professor Emeritus
Committee Member, Department of Computer Science

K. Bradley Paxton, Ph.D., CEO
Committee Member, ADI, LLC

Thesis Release Permission Form

Rochester Institute of Technology
B. Thomas Golisano College of Computing and Information Sciences

Title:

Balancing Truth Error and Manual Processing in the PDQ System

I, Douglass Huang, hereby grant permission to the Wallace Memorial Library to reproduce my thesis in whole or part.

Douglass Huang

Date

Dedication

To Nina, Eric, Peter, Jaki, and Sir Duke.

Acknowledgments

I am grateful to the following people for all their support, patience, and encouragement: my wife, Mrs. Nina Huang; my parents, Dr. Ji-Ming Huang and Mrs. Chuei-Yun Huang; my committee members, Dr. Roger Gaborski, Dr. Peter Anderson, and Dr. Brad Paxton; Graduate Program Coordinator, Dr. Hans-Peter Bischof; my boss, Mr. Steve Spiwak; my current and past co-workers at ADI, LLC; and all my family and friends.

Abstract

Balancing Truth Error and Manual Processing in the PDQ System

Douglass Huang

Supervising Professor: Roger S. Gaborski, Ph.D.

Production Data Quality (PDQ) is a specialized pattern classifier whose main purpose is to assess independently the data quality of a production classifier. It accomplishes this by producing a high quality Truth from the source input, and then using the Truth to identify errors in the production classifier's output data. Previous studies have shown close agreement between PDQ processing outcomes and a particular mathematical model of the system.

In this study we describe and analyze an expanded model that addresses the potential tradeoff between Truth error and manual processing in PDQ, with an eye towards informing operational decisions about precision and efficiency. Using statistical data from the 2010 Census PDQ system as input, we examine the predictions of the new model in order to understand their potential usefulness.

The outcomes show strong agreement between two methods for estimating Projected Truth error rate, supporting the validity of both methods as well as the existing static model. In addition, the new Projector model gives tight bounds on the projected manual processing rate and reveals a characteristic relationship between Projected Truth error and projected manual processing. We explore a practical application of this model for tuning PDQ, and we find an opportunity to achieve a 60% efficiency increase for the selected sample, while maintaining an acceptable degree of precision.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Production Data Quality (PDQ) [1] is a system developed at ADI, LLC to measure the accuracy of a pattern classifier when processing "production" inputs, where the true classifications are not known *a priori*. It can be applied generically in a number of classification domains, such as fingerprint matching or record linkage. In one particular instance, PDQ has been used to assess the quality of the Decennial Response Integration System's (DRIS) [13] electronic capture of handprint and check mark responses on 2010 Census paper forms. In order to make its measurements, PDQ first produced a high quality Truth for a sample of scanned Census form images, using a combination of automated recognition and human processing. Paxton, *et al.*, have described a mathematical model [12] to predict the outcomes of this process based on certain input conditions, and in practice the actual outcomes have agreed very well with the predictions.

PDQ itself is a special case of pattern classifier, and as such, the simplest measure of its performance is the *Truth error rate*, the fraction of response fields for which it assigns an incorrect Truth value [6]. Because its main purpose is to measure the accuracy of a production classifier system, PDQ's Truth error rate must be sufficiently low in comparison to the error rate of the production output data. Another useful measure of PDQ's performance is the *manual processing rate*, expressed as the fraction of fields that require human review or arbitration to determine the Truth. Reducing this workload can result in reduced labor costs or increased throughput, but likely additional Truth error.

This study examines the potential tradeoff between Truth error and manual processing in PDQ, which the previous, static model does not address.

To that end we define an expanded model to predict the impact on these two performance measures, given a change in the *confidence threshold*, an operating parameter of PDQ's Independent Data Capture system component. Using historical processing results from the 2010 Census, we explore and analyze the various predictions of this new Projector model, and we see how the model can be applied practically for making optimal operational decisions.

Chapter 2

Supporting Work

2.1 Production Data Quality

2.1.1 Process Flow

In its 2010 Census embodiment, the PDQ system has the following main components [12, 7], shown in Figure 2.1:

Independent Data Capture system Processes the Census form images and outputs the Provisional Truth (data set PT). This is analogous to the Production Data Capture system, whose Production Data (data set PD) is being evaluated, but it has been developed independently to meet equivalent input/output specifications.

Comparator 1 (C1) Determines automatically, on a field-by-field basis, whether the response values in the Production Data and Provisional Truth are identical (path PT=PD). Matching values are designated as Truth (data set T) and require no further processing.

Arbitrator Incorporates human analysts to resolve the Truth for any non-matching fields (path PT=Other1) identified by Comparator 1.

2.1.2 Arbitrator

The steps within the Arbitrator, shown in Figure 2.2, are as follows:

Analyst 1 Enters a value for each field sent to the Arbitrator, producing the data set A1.

Comparator 2 (C2) Determines automatically whether Analyst 1's value matches that of the Production Data (path A1=PD) or Provisional Truth (path A1=PT). If so, that value is designated as Truth and requires no further processing.

Analyst 2 Enters a value for each non-matching field (A1=Other2) identified by C2, producing the data set A2.

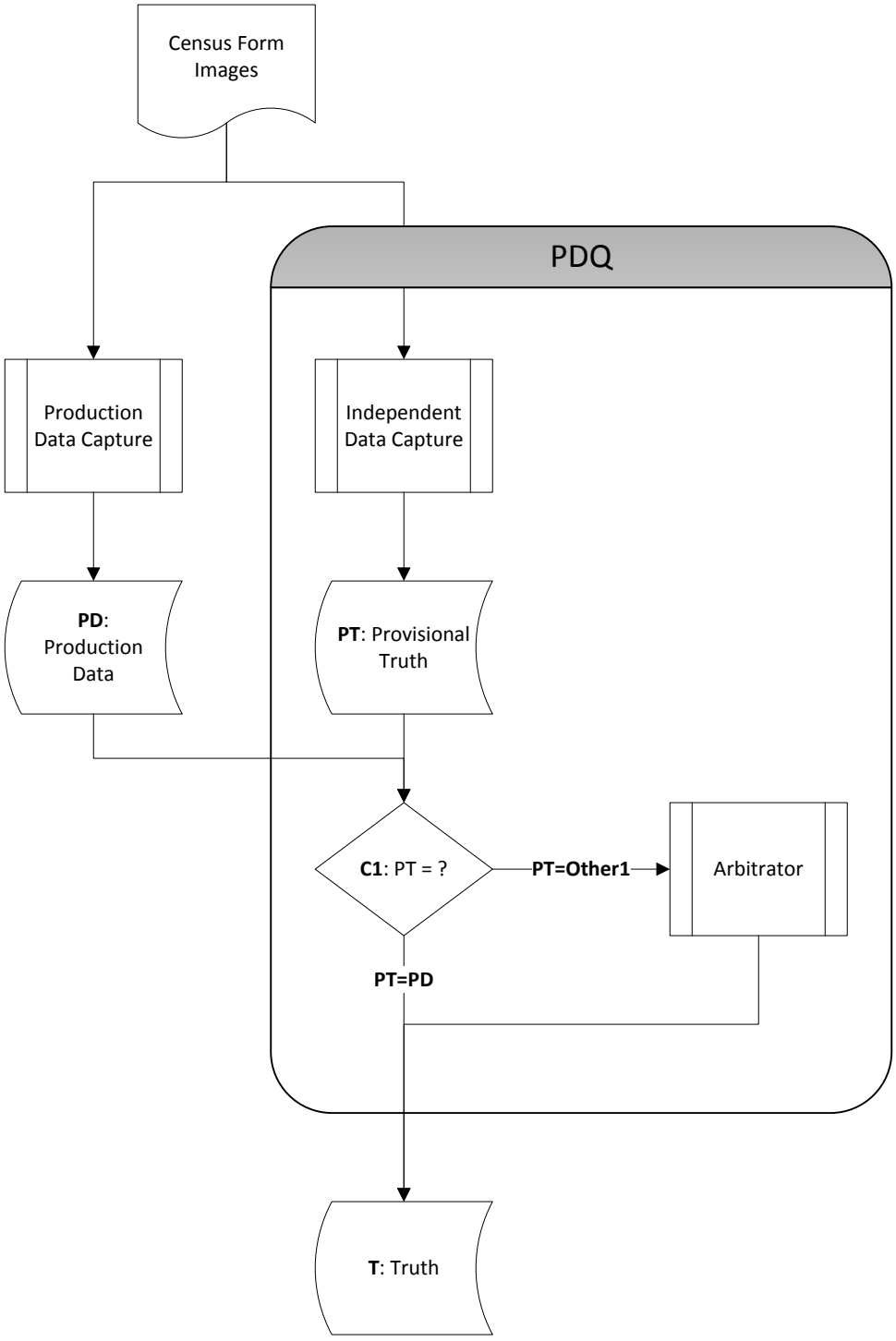


Figure 2.1: PDQ Process Flow: Main components.

Comparator 3 (C3) Determines automatically whether Analyst 2's value matches that of the Production Data (path A2=PD), Provisional Truth (path A2=PT), or Analyst 1 (path A2=A1). If so, that value is designated as Truth. Otherwise (path A2=Other3, or INC), PDQ designates the field as Inconclusive.

2.1.3 Static Model

The mathematical model of PDQ given by Paxton, *et al.* [12], takes as inputs the *error rates* of four of the data sets identified above. Error rates are defined as follows:

$$\begin{aligned} f_x &\equiv \text{the number of fields in data set } x \\ e_x &\equiv \text{the number of errors in data set } x \\ E_x &\equiv \text{the error rate of data set } x \\ E_x &= \frac{e_x}{f_x} \end{aligned} \tag{2.1}$$

The inputs to the model are the following:

$$\begin{aligned} E_{\text{PD}} &\equiv \text{the error rate of the Production Data} \\ E_{\text{PT}} &\equiv \text{the error rate of the Provisional Truth} \\ E_{\text{A1}} &\equiv \text{the error rate of Analyst 1's output} \\ E_{\text{A2}} &\equiv \text{the error rate of Analyst 2's output} \end{aligned}$$

We define the set Y as follows:

$$\begin{aligned} Y &\equiv \text{the set of all direct paths } y \text{ to the Truth} \\ Y &= \{\text{PT=PD, A1=PD, A1=PT, A2=PD, A2=PT, A2=A1, INC}\} \end{aligned} \tag{2.2}$$

Assuming that the inputs are independent random variables [8], the model

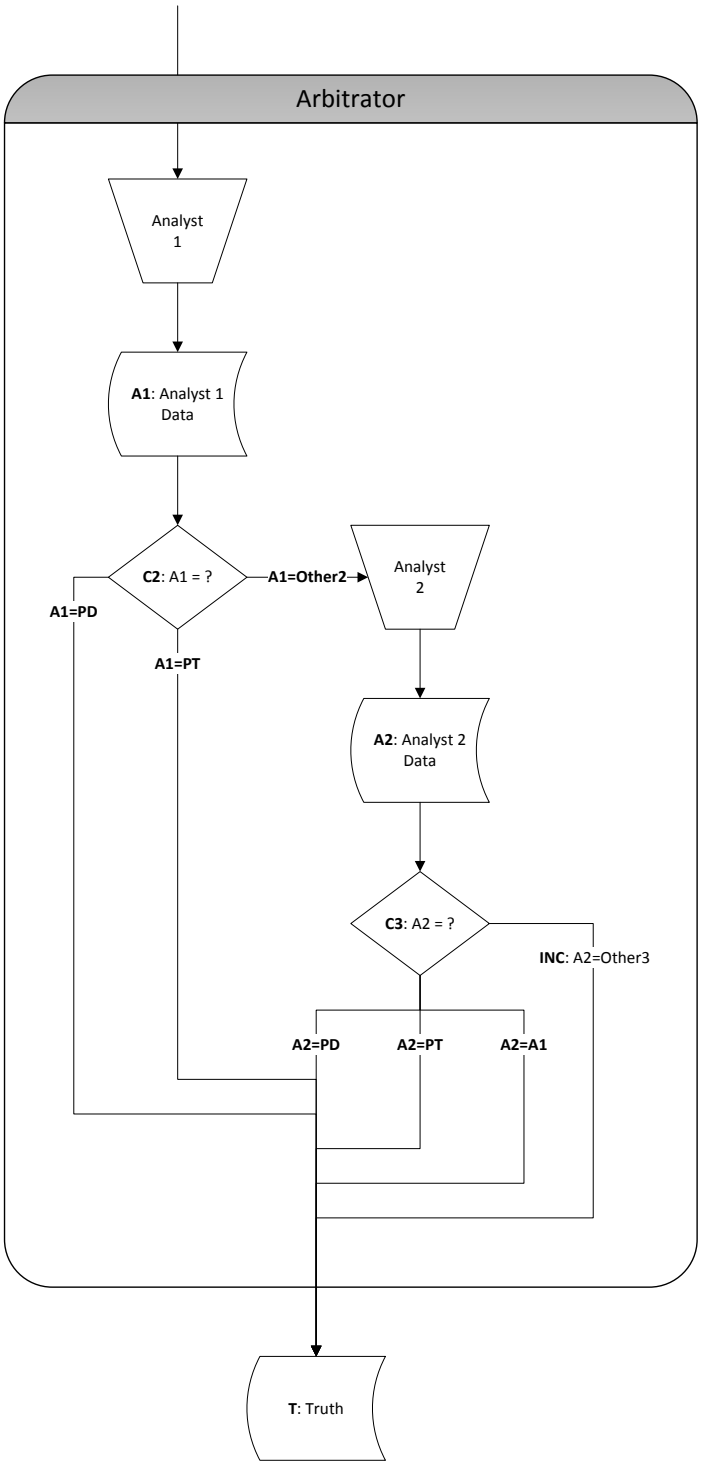


Figure 2.2: PDQ Process Flow: Arbitrator steps.

predicts the outcomes below:

$$P[y] \equiv \text{the probability that a field follows path } y$$

$$P[\text{PT=PD}] = (1 - E_{\text{PD}})(1 - E_{\text{PT}}) \tag{2.3a}$$

$$P[\text{A1=PD}] = (1 - E_{\text{PD}})E_{\text{PT}}(1 - E_{\text{A1}}) \tag{2.3b}$$

$$P[\text{A1=PT}] = E_{\text{PD}}(1 - E_{\text{PT}})(1 - E_{\text{A1}}) \tag{2.3c}$$

$$P[\text{A2=PD}] = (1 - E_{\text{PD}})E_{\text{PT}}E_{\text{A1}}(1 - E_{\text{A2}}) \tag{2.3d}$$

$$P[\text{A2=PT}] = E_{\text{PD}}(1 - E_{\text{PT}})E_{\text{A1}}(1 - E_{\text{A2}}) \tag{2.3e}$$

$$P[\text{A2=A1}] = E_{\text{PD}}E_{\text{PT}}(1 - E_{\text{A1}})(1 - E_{\text{A2}}) \tag{2.3f}$$

$$P[\text{INC}] = E_{\text{PD}}E_{\text{PT}}(E_{\text{A1}} + E_{\text{A2}}) + E_{\text{A1}}E_{\text{A2}}(E_{\text{PD}} + E_{\text{PT}}) - 3E_{\text{PD}}E_{\text{PT}}E_{\text{A1}}E_{\text{A2}} \tag{2.3g}$$

$$\sum_{y \in Y} P[y] = 1 \tag{2.3h}$$

2.1.4 Truth Error Rate

We define $f[y]$ and $F[y]$ as follows:

$$f[y] \equiv \text{the number of fields that follow path } y$$

$$F[y] \equiv \text{the rate at which fields follow path } y$$

$$F[y] = \frac{f[y]}{f_{\text{T}}} \tag{2.4}$$

Paxton [11] gives the following estimate of the PDQ Truth error rate E_{T}:

$$E_{\text{T}} = \max_{y \in Y} |F[y] - P[y]| \tag{2.5}$$

For the purposes of my study, I assume that this is a practical baseline estimate. In one representative subset of 2010 Census results [7], PDQ measured a Production Data error rate of $E_{\text{PD}} = 0.28\%$, while the Truth error rate given by Equation (2.5) was a substantially lower $E_{\text{T}} = 0.01\%$, indicating strong agreement between the modeled and actual outcomes [12]. As discussed later, my analysis focuses on the estimated *change* in Truth error rate that results from changing an operating parameter of PDQ's Independent Data Capture system.

2.1.5 Independent Data Capture System

Next, we take a more detailed look at the following steps within the Independent Data Capture system [12], shown in Figure 2.3:

Recognizer Uses OCR or optical mark recognition (OMR) to assign automatically a response value and an integral confidence level in the range $[0, 100]$ for each field on each Census form image, producing the data set R.

Acceptor 1 (AR1) Determines automatically, based on the Recognizer outputs and an integral confidence threshold $\theta_0 \in [-1, 100]$, whether the value of each field in data set R is *accepted*. While there are specific exceptions due to complex contextual rules, in general a field is accepted if its confidence level *exceeds* θ_0. If it is accepted (path ACC), the field contributes to the data set $\mathrm{R_{ACC}}$, which becomes part of the Provisional Truth.

Keyer Performs manual review of each field *rejected* by the Acceptor (path REJ). Enters a value into the data set K, which completes the Provisional Truth data set.

2.1.6 Manual Processing Rate

While the E_{PT} input of the static model reflects one performance measure of the Independent Data Capture system, it does not account for the *reject rate*, which reflects the amount of human review required to produce the Provisional Truth. We define the reject rate F_{K} as follows:

$$F_x \equiv \text{the rate at which fields contribute to data set } x$$

$$F_x = \frac{f_x}{f_{\mathrm{T}}} \tag{2.6}$$

$$F_{\mathrm{K}} = \text{the reject rate of the Independent Data Capture system}$$

$$F_{\mathrm{K}} = \frac{f_{\mathrm{K}}}{f_{\mathrm{T}}} \tag{2.7}$$

An analogous performance measure for PDQ as a whole is the *manual processing rate* [11], which reflects the total amount of human review and arbitration required to determine the Truth. We define the manual processing rate M_T as follows:

$$m_{\mathrm{T}} \equiv \text{the manual processing volume required to determine the Truth}$$

$$m_{\mathrm{T}} = f_{\mathrm{K}} + f_{\mathrm{A1}} + f_{\mathrm{A2}} \tag{2.8}$$

$$M_{\mathrm{T}} \equiv \text{the manual processing rate required to determine the Truth}$$

$$M_{\mathrm{T}} = \frac{m_{\mathrm{T}}}{f_{\mathrm{T}}} \tag{2.9}$$

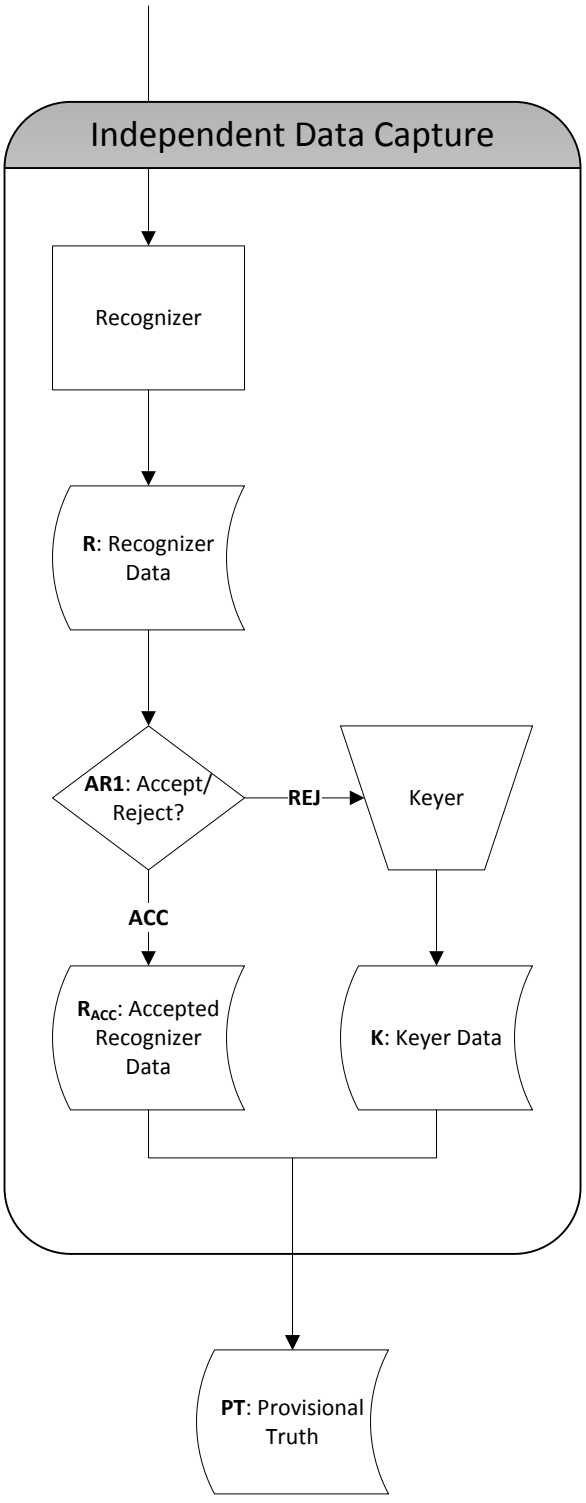


Figure 2.3: PDQ Process Flow: Independent Data Capture steps.

2.2 Error and Manual Processing Tradeoff

The tradeoff between error and reject has been well studied as a means to characterize and optimize the performance of handwriting recognition systems [3, 2]. A typical approach begins with processing a training deck of known truth using an automated recognizer. The recognizer outputs both a response value and an integral confidence level for each work unit. One can then determine the proportion of fields whose confidence levels are below different possible confidence thresholds (reject rate), and also the proportion of incorrect response values among accepted fields (error rate). These kinds of data, especially in conjunction with a cost model [4, 9], can be used to determine optimal confidence thresholds for the system.

The most obvious application in PDQ is to examine how the confidence threshold θ_0 within the Independent Data Capture system impacts both the Provisional Truth error rate and the reject rate. It also follows that the choice of confidence threshold impacts the overall Truth error rate and manual processing rate in PDQ. My study examines these relationships in detail.

Chapter 3

Expanded Model of PDQ Outcomes

PDQ's main role in the 2010 Census was to verify the DRIS contractor's adherence to certain data quality service-level agreements (SLAs), expressed in terms of the error rate E_{PD}, for specific strata within the Production Data set. For example, the total write-in fields captured by optical character recognition (OCR) with high confidence were required to have $E_{\text{PD}} \leq 1.0\%$. It stands to reason that a certain Truth error rate, perhaps $E_{\text{T}} = 0.1\%$, would be sufficiently low for verifying SLA compliance, while an even lower Truth error rate would be unnecessary for that purpose. Because lower Truth error rates typically come at the cost of additional manual processing, it would be helpful to have some control over this tradeoff. I have devised an expanded view of the PDQ processing outcomes that attempts to address this concern.

3.1 Projector

Given some estimate of E_{T} and an observed M_{T}, we wish to predict how each of these performance measures would change after reducing the confidence threshold from θ_0. For this purpose I introduce the Projector, a logical component that examines the processing outcomes (*i.e.*, data set contents and comparator decisions) from PDQ and computes the projected outcomes for integral confidence thresholds θ, where $\theta \in [-1, \theta_0]$. The case $\theta = -1$ results in acceptance of the most fields possible, while $\theta = \theta_0$ results in the original PDQ outcomes. Figure 3.1 shows the steps within the Projector, which are as follows:

Path Identifier 1 (PI1) Inspects automatically which path $y \in Y$ each field followed originally in PDQ.

Path Identifier 2 (PI2) Inspects automatically the decision of Acceptor 1. Fields that were accepted (path ACC) at the original confidence threshold θ_0 would be unchanged by applying a lower threshold. Fields that were rejected (path REJ) require further consideration.

Acceptor 2 (AR2(θ)) Determines automatically whether each field that was previously rejected would be accepted at the new threshold θ. If there would be no change (path REJ(θ)), then no further analysis is needed. For fields that would be accepted (path ACC(θ)), the original Recognizer values are added to the data set $\mathrm{R_{ACC}}(\theta)$.

Comparator 4 (C4) Determines automatically whether the value for each field in $\mathrm{R_{ACC}}(\theta)$ matches that from either the Production Data (path $\mathrm{R_{ACC}}(\theta)$=PD), original Provisional Truth (path $\mathrm{R_{ACC}}(\theta)$=PT), Analyst 1 (path $\mathrm{R_{ACC}}(\theta)$=A1), Analyst 2 (path $\mathrm{R_{ACC}}(\theta)$=A2), or none of these (path $\mathrm{R_{ACC}}(\theta)$=Other4).

Path Predictor Determines automatically the Projected Truth (data set T(θ)) by predicting a new terminal path $y(\theta)$ based on y and the incoming path $z(\theta)$ from PI2, AR2(θ), or C4. In certain cases, $y(\theta)$ is indeterminate, as described later in Section 3.2.

3.2 Path Predictor

To support the description of the Path Predictor's function, we define the following:

$$Z(\theta) \equiv \text{the set of all direct paths } z(\theta) \text{ to the Path Predictor}$$
$$Z(\theta) = \{\mathrm{R_{ACC}}(\theta)\text{=PD}, \mathrm{R_{ACC}}(\theta)\text{=PT}, \mathrm{R_{ACC}}(\theta)\text{=A1}, \mathrm{R_{ACC}}(\theta)\text{=A2}, \mathrm{R_{ACC}}(\theta)\text{=Other4}\} \tag{3.1}$$

$$Y(\theta) \equiv \text{the set of all direct paths } y(\theta) \text{ to the Projected Truth}$$
$$Y(\theta) = Y \cup \{\text{PT}(\theta)\text{=PD}, \text{A1=PT}(\theta), \text{A2=PT}(\theta)\} \tag{3.2}$$
$$Y(\theta)[y, z(\theta)] \equiv \text{the set of paths } y(\theta) \text{ that are possible for each field that follows the paths } y \text{ and } z(\theta)$$

Table 3.1 shows the sets of possible paths $Y(\theta)[y, z(\theta)]$ that the Path Predictor computes from the various combinations of y and $z(\theta)$. The mappings are derived by substituting the $\mathrm{R_{ACC}}(\theta)$ values for the original Provisional Truth values of applicable fields, and then "replaying" the PDQ process flow using this Projected Provisional Truth (data set PT(θ)) and the other original data sets. The operative assumption is that pre-existing

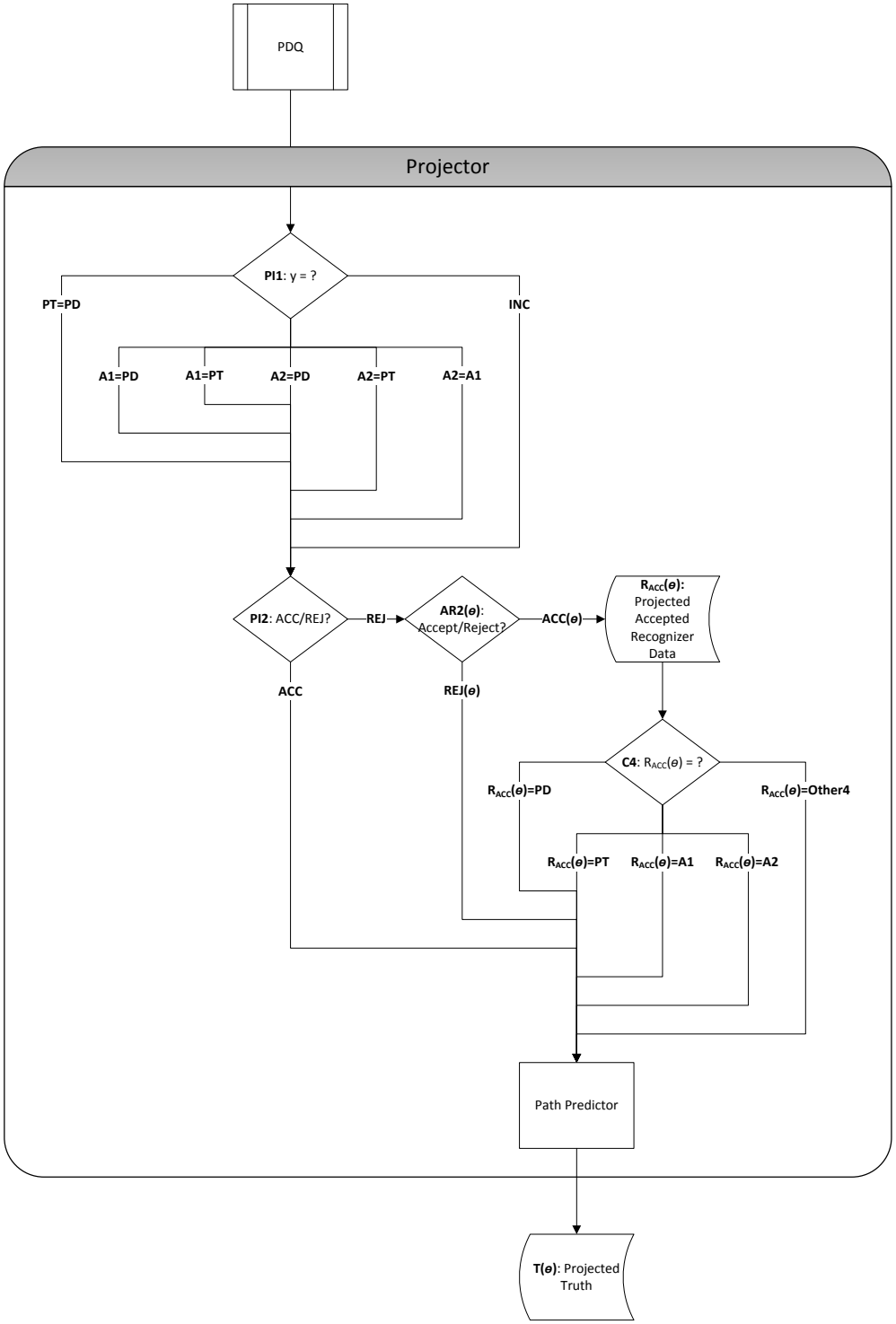


Figure 3.1: Expanded PDQ Model: Projector steps.

Table 3.1: Path Predictor: $Y(\theta)[y, z(\theta)]$

y	$z(\theta)$: ACC	REJ(θ)	$R_{ACC}(\theta)$=PD	$R_{ACC}(\theta)$=PT	$R_{ACC}(\theta)$=A1	$R_{ACC}(\theta)$=A2	$R_{ACC}(\theta)$=Other4
PT=PD	{PT=PD}	{PT=PD}	{PT=PD}	$\emptyset$	$\emptyset$	$\emptyset$	{A1=PD, A1=PT(θ), A2=PD, A2=PT(θ), A2=A1, INC}
A1=PD	{A1=PD}	{A1=PD}	{PT(θ)=PD}	{A1=PD}	$\emptyset$	$\emptyset$	{A1=PD}
A1=PT	{A1=PT}	{A1=PT}	{PT(θ)=PD}	{A1=PT}	$\emptyset$	$\emptyset$	{A2=PD, A2=PT(θ), A2=A1, INC}
A2=PD	{A2=PD}	{A2=PD}	{PT(θ)=PD}	{A2=PD}	{A1=PT(θ)}	$\emptyset$	{A2=PD}
A2=PT	{A2=PT}	{A2=PT}	{PT(θ)=PD}	{A2=PT}	{A1=PT(θ)}	$\emptyset$	{INC}
A2=A1	{A2=A1}	{A2=A1}	{PT(θ)=PD}	{A2=A1}	{A1=PT(θ)}	$\emptyset$	{A2=A1}
INC	{INC}	{INC}	{PT(θ)=PD}	{INC}	{A1=PT(θ)}	{A2=PT(θ)}	{INC}

values in data sets PD, A1, and A2 would remain unchanged in the projection. Note that certain outcomes are indeterminate (*e.g.*, in the case $(y, z(\theta)) = (\text{PT=PD}, R_{ACC}(\theta)\text{=Other4})$ due to fields having bypassed one or both Analyst steps originally. Certain other combinations of y and $z(\theta)$ are invalid in the Projector's process flow, so they are shown to have no possible outcomes.

The following examples illustrate the logic encapsulated in the Path Predictor, as shown in Figure 3.1 and Table 3.1:

- Regardless of the original terminal path y, if PI2 identifies a given field as having been accepted by AR1 ($z(\theta) = \text{ACC}$), then the Provisional Truth value is unchanged under the new threshold θ, and there is no change in the field's terminal path ($Y(\theta)[y, \text{ACC}] = \{y\}$ for all y).

- If a given field's original terminal path y is PT=PD, PI2 identifies it as having been rejected by AR1 (path REJ), AR2(θ) accepts it (path ACC(θ)), and C4 determines that its Recognizer value matches the Production Data value ($z(\theta) = \langle R_{ACC}(\theta)\text{=PD}\rangle$); then its projected terminal path is unchanged ($y(\theta) = \langle\text{PT=PD}\rangle$).

- In the case where $y = \langle\text{A1=PT}\rangle$ and the Recognizer value matches neither the Production Data value nor the original Provisional Truth

value ($z(\theta) = \langle \mathrm{R_{ACC}}(\theta)\text{=Other4} \rangle$), then in the projection the field must be sent to Analyst 2. Since there is no original A2 value, four projected terminal paths $y(\theta)$ are possible ($Y(\theta)[\text{A1=PT}, \mathrm{R_{ACC}}(\theta)\text{=Other4}] = \{\text{A2=PD}, \text{A2=PT}(\theta), \text{A2=A1}, \text{INC}\}$).

It will be useful in later discussions to account separately for the determinate and indeterminate cases. For that purpose, we define the following sets:

$YZ(\theta) \equiv$ the set of all pairs $(y, z(\theta))$

$$YZ(\theta) = Y \times Z(\theta) \tag{3.3}$$

$YZ(\theta)_0 \equiv$ the set of all invalid pairs $(y, z(\theta))$

$$YZ(\theta)_0 = \{(y, z(\theta)) \in YZ(\theta) : |Y(\theta)[y, z(\theta)]| = 0\} \tag{3.4}$$

$YZ(\theta)_1 \equiv$ the set of all valid pairs $(y, z(\theta))$ for which the resultant path $y(\theta)$ is determinate

$$YZ(\theta)_1 = \{(y, z(\theta)) \in YZ(\theta) : |Y(\theta)[y, z(\theta)]| = 1\} \tag{3.5}$$

$YZ(\theta)_2 \equiv$ the set of all valid pairs $(y, z(\theta))$ for which the resultant path $y(\theta)$ is indeterminate

$$YZ(\theta)_2 = \{(y, z(\theta)) \in YZ(\theta) : |Y(\theta)[y, z(\theta)]| \geq 2\} \tag{3.6}$$

$$YZ(\theta)_2 = \{(\text{PT=PD}, \mathrm{R_{ACC}}(\theta)\text{=Other4}), (\text{A1=PT}, \mathrm{R_{ACC}}(\theta)\text{=Other4})\} \tag{3.7}$$

These sets have the following additional properties:

$$YZ(\theta) = YZ(\theta)_0 \cup YZ(\theta)_1 \cup YZ(\theta)_2 \tag{3.8}$$

$$|YZ(\theta)| = |YZ(\theta)_0| + |YZ(\theta)_1| + |YZ(\theta)_2| \tag{3.9}$$

That is, the three subsets $YZ(\theta)_0$, $YZ(\theta)_1$, $YZ(\theta)_2$ are pairwise disjoint, and together they comprise the complete set $YZ(\theta)$.

3.3 Projected Truth Error Rate: Method 1

If we assume that any change from the Truth — a difference in INC status, or otherwise a different value in the Projected Truth $\mathrm{T}(\theta)$ — constitutes an additional Projected Truth error, then we can determine the incremental change in Projected Truth error count for each field, given the paths y, $z(\theta)$,

and $y(\theta)$. We define the function $\delta e_{\mathrm{T}(\theta)}[y, z(\theta), y(\theta)]$ as follows:

$$\delta e_{\mathrm{T}(\theta)}[y, z(\theta), y(\theta)] \equiv \text{the incremental change in Projected Truth error count for each field that follows the paths } y, z(\theta), \text{ and } y(\theta)$$

$$\delta e_{\mathrm{T}(\theta)}[y, z(\theta), y(\theta)] = \begin{cases} 1, & \text{if } y = \langle x_1{=}x_2 \rangle \text{ and } y(\theta) = \langle x_3{=}x_4 \rangle \text{ and } \{x_1, x_2\} \cap \{x_3, x_4\} = \emptyset \\ 1, & \text{if } y = \text{INC and } y(\theta) \neq \text{INC} \\ 1, & \text{if } y \neq \text{INC and } y(\theta) = \text{INC} \\ 0, & \text{otherwise} \end{cases} \tag{3.10}$$

Because $y(\theta)$ is indeterminate in some cases, we define a minimum and maximum incremental change in Projected Truth error count, given the paths y and $z(\theta)$.

$$\min \delta e_{\mathrm{T}(\theta)}[y, z(\theta)] \equiv \text{the minimum incremental change in Projected Truth error count for each field that follows the paths } y \text{ and } z(\theta)$$

$$\min \delta e_{\mathrm{T}(\theta)}[y, z(\theta)] = \begin{cases} 1, & \text{if } (y, z(\theta)) \in YZ(\theta)_1 \text{ and } y = \langle x_1{=}x_2 \rangle \text{ and } Y(\theta)[y, z(\theta)] = \{x_3{=}x_4\} \\ & \quad \text{and } \{x_1, x_2\} \cap \{x_3, x_4\} = \emptyset \\ 1, & \text{if } (y, z(\theta)) \in YZ(\theta)_1 \text{ and } y = \text{INC and } Y(\theta)[y, z(\theta)] \neq \{\text{INC}\} \\ 1, & \text{if } (y, z(\theta)) \in YZ(\theta)_1 \text{ and } y \neq \text{INC and } Y(\theta)[y, z(\theta)] = \{\text{INC}\} \\ 0, & \text{if } (y, z(\theta)) = (\text{PT=PD}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4}) \\ 0, & \text{if } (y, z(\theta)) = (\text{A1=PT}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4}) \\ 0, & \text{otherwise} \end{cases} \tag{3.11}$$

$$\max \delta e_{\mathrm{T}(\theta)}[y, z(\theta)] \equiv \text{the maximum incremental change in Projected Truth error count for each field that follows the paths } y \text{ and } z(\theta)$$

$$\max \delta e_{\mathrm{T}(\theta)}[y, z(\theta)] = \begin{cases} 1, & \text{if } (y, z(\theta)) \in YZ(\theta)_1 \text{ and } y = \langle x_1{=}x_2 \rangle \text{ and } Y(\theta)[y, z(\theta)] = \{x_3{=}x_4\} \\ & \quad \text{and } \{x_1, x_2\} \cap \{x_3, x_4\} = \emptyset \\ 1, & \text{if } (y, z(\theta)) \in YZ(\theta)_1 \text{ and } y = \text{INC and } Y(\theta)[y, z(\theta)] \neq \{\text{INC}\} \\ 1, & \text{if } (y, z(\theta)) \in YZ(\theta)_1 \text{ and } y \neq \text{INC and } Y(\theta)[y, z(\theta)] = \{\text{INC}\} \\ 1, & \text{if } (y, z(\theta)) = (\text{PT=PD}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4}) \\ 1, & \text{if } (y, z(\theta)) = (\text{A1=PT}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4}) \\ 0, & \text{otherwise} \end{cases} \tag{3.12}$$

We will examine the change in Projected Truth error count for fields that follow each path $z(\theta)$, as follows:

$$\min \Delta e_{\mathrm{T}(\theta)}[z(\theta)] \equiv \text{the minimum total change in Projected Truth error count for all fields that follow the path } z(\theta)$$

$$\min \Delta e_{\mathrm{T}(\theta)}[z(\theta)] = \sum_{y \in Y} \min \delta e_{\mathrm{T}(\theta)}[y, z(\theta)] f[y, z(\theta)] \tag{3.13}$$

$$\max \Delta e_{\mathrm{T}(\theta)}[z(\theta)] \equiv \text{the maximum total change in Projected Truth error count for all fields that follow the path } z(\theta)$$

$$\max \Delta e_{\mathrm{T}(\theta)}[z(\theta)] = \sum_{y \in Y} \max \delta e_{\mathrm{T}(\theta)}[y, z(\theta)] f[y, z(\theta)] \tag{3.14}$$

There are four $z(\theta)$ paths that can contribute additional Projected Truth errors, and the following are true for them:

$$\max \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=PD}] = \min \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=PD}] \tag{3.15}$$

$$\max \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A1}] = \min \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A1}] \tag{3.16}$$

$$\max \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A2}] = \min \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A2}] \tag{3.17}$$

$$\begin{aligned}\max \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=Other4}] = {} & \min \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=Other4}] + f[\text{PT=PD}, \mathrm{R_{ACC}}(\theta)\text{=Other4}] \\ & + f[\text{A1=PT}, \mathrm{R_{ACC}}(\theta)\text{=Other4}]\end{aligned} \tag{3.18}$$

Dropping the $\min$ and $\max$ designations as appropriate, we can compute the total change in Projected Truth error count as follows:

$$\min \Delta e_{\mathrm{T}(\theta)} \equiv \text{the minimum total change in Projected Truth error count}$$

$$\begin{aligned}\min \Delta e_{\mathrm{T}(\theta)} = {} & \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=PD}] + \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A1}] + \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A2}] \\ & + \min \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=Other4}]\end{aligned} \tag{3.19}$$

$$\max \Delta e_{\mathrm{T}(\theta)} \equiv \text{the maximum total change in Projected Truth error count}$$

$$\begin{aligned}\max \Delta e_{\mathrm{T}(\theta)} = {} & \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=PD}] + \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A1}] + \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A2}] \\ & + \max \Delta e_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=Other4}]\end{aligned} \tag{3.20}$$

$$\max \Delta e_{\mathrm{T}(\theta)} = \min \Delta e_{\mathrm{T}(\theta)} + f[\text{PT=PD}, \mathrm{R_{ACC}}(\theta)\text{=Other4}] + f[\text{A1=PT}, \mathrm{R_{ACC}}(\theta)\text{=Other4}] \tag{3.21}$$

Then, we divide to obtain the total change in Projected Truth error rate:

$$\min \Delta E_{\mathrm{T}(\theta)} \equiv \text{the minimum total change in Projected Truth error rate}$$

$$\min \Delta E_{\mathrm{T}(\theta)} = \frac{\min \Delta e_{\mathrm{T}(\theta)}}{f_{\mathrm{T}}} \tag{3.22}$$

$$\min \Delta E_{\mathrm{T}(\theta)} = \Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=PD}] + \Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A1}] + \Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A2}] + \min \Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=Other4}] \tag{3.23}$$

$$\max \Delta E_{\mathrm{T}(\theta)} \equiv \text{the maximum total change in Projected Truth error rate}$$

$$\max \Delta E_{\mathrm{T}(\theta)} = \frac{\max \Delta e_{\mathrm{T}(\theta)}}{f_{\mathrm{T}}} \tag{3.24}$$

$$\max \Delta E_{\mathrm{T}(\theta)} = \min \Delta E_{\mathrm{T}(\theta)} + F[\text{PT=PD}, \mathrm{R_{ACC}}(\theta)\text{=Other4}] + F[\text{A1=PT}, \mathrm{R_{ACC}}(\theta)\text{=Other4}] \tag{3.25}$$

Thus we compute the total Projected Truth error rate by Method 1 as follows:

$$\min E1_{\mathrm{T}(\theta)} = E_{\mathrm{T}} + \min \Delta E_{\mathrm{T}(\theta)} \tag{3.26}$$

$$\min E1_{\mathrm{T}(\theta)} = E_{\mathrm{T}} + \Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=PD}] + \Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A1}] + \Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A2}] + \min \Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=Other4}] \tag{3.27}$$

$$\max E1_{\mathrm{T}(\theta)} = E_{\mathrm{T}} + \max \Delta E_{\mathrm{T}(\theta)} \tag{3.28}$$

$$\max E1_{\mathrm{T}(\theta)} = \min E1_{\mathrm{T}(\theta)} + F[\text{PT=PD}, \mathrm{R_{ACC}}(\theta)\text{=Other4}] + F[\text{A1=PT}, \mathrm{R_{ACC}}(\theta)\text{=Other4}] \tag{3.29}$$

3.4 Projected Provisional Truth Error Rate

Next, we will examine how the Projected Provisional Truth error rate $E_{\text{PT}(\theta)}$ is computed. We define the following functions:

$\varepsilon e_{\text{R}_{\text{ACC}}(\theta)}[y, z(\theta)] \equiv$ the incremental contribution to Recognizer-accepted error count in the Projected Provisional Truth for each field that follows the paths y and $z(\theta)$

$$\varepsilon e_{\text{R}_{\text{ACC}}(\theta)}[y, z(\theta)] = \begin{cases} 1, & \text{if } y = \langle x_1{=}x_2 \rangle \text{ and } z(\theta) = \text{ACC and PT} \notin \{x_1, x_2\} \\ 1, & \text{if } y = \langle x_1{=}x_2 \rangle \text{ and } z(\theta) = \langle \text{R}_{\text{ACC}}(\theta){=}x_3 \rangle \text{ and } x_3 \notin \{x_1, x_2\} \\ 1, & \text{if } y = \text{INC and } z(\theta) \neq \text{REJ}(\theta) \\ 0, & \text{otherwise} \end{cases} \tag{3.30}$$

$\varepsilon e_{\text{K}(\theta)}[y, z(\theta)] \equiv$ the incremental contribution to Keyer error count in the Projected Provisional Truth for each field that follows the paths y and $z(\theta)$

$$\varepsilon e_{\text{K}(\theta)}[y, z(\theta)] = \begin{cases} 1, & \text{if } y = \langle x_1{=}x_2 \rangle \text{ and } z(\theta) = \text{REJ}(\theta) \text{ and PT} \notin \{x_1, x_2\} \\ 1, & \text{if } y = \text{INC and } z(\theta) = \text{REJ}(\theta) \\ 0, & \text{otherwise} \end{cases} \tag{3.31}$$

The total error count in each data set is as follows:

$$e_{\text{R}_{\text{ACC}}(\theta)}[y, z(\theta)] = \varepsilon e_{\text{R}_{\text{ACC}}(\theta)}[y, z(\theta)] f_{\text{R}_{\text{ACC}}(\theta)}[y, z(\theta)] \tag{3.32}$$

$$e_{\text{K}(\theta)}[y, z(\theta)] = \varepsilon e_{\text{K}(\theta)}[y, z(\theta)] f_{\text{K}(\theta)}[y, z(\theta)] \tag{3.33}$$

We can then compute the Projected Provisional Truth error count and error rate:

$$e_{\text{R}_{\text{ACC}}(\theta)} = \sum_{y \in Y} \sum_{z \in Z(\theta)} e_{\text{R}_{\text{ACC}}(\theta)}[y, z(\theta)] \tag{3.34}$$

$$e_{\text{K}(\theta)} = \sum_{y \in Y} \sum_{z \in Z(\theta)} e_{\text{K}(\theta)}[y, z(\theta)] \tag{3.35}$$

$$e_{\text{PT}(\theta)} = e_{\text{R}_{\text{ACC}}(\theta)} + e_{\text{K}(\theta)} \tag{3.36}$$

$$E_{\text{PT}(\theta)} = \frac{e_{\text{PT}(\theta)}}{f_{\text{T}}} \tag{3.37}$$

$$E_{\text{PT}(\theta)} = E_{\text{R}_{\text{ACC}}(\theta)} + E_{\text{K}(\theta)} \tag{3.38}$$

3.5 Projected Truth Error Rate: Method 2

Given the Projected Provisional Truth error rate, we have another method for computing the Projected Truth error rate, by substituting $E_{\text{PT}(\theta)}$ for E_{PT} as an input to the static model.

First, because $Y \subset Y(\theta)$, we define the path-equivalent $y'(\theta) \in Y$ as follows:

$$
\begin{aligned}
y'(\theta) &\equiv \text{the path-equivalent in } Y \text{ of the path } y(\theta) \\
y'(\theta) &= \begin{cases} \text{PT=PD}, & \text{if } y(\theta) \in \{\text{PT=PD}, \text{PT}(\theta)\text{=PD}\} \\ \text{A1=PT}, & \text{if } y(\theta) \in \{\text{A1=PT}, \text{A1=PT}(\theta)\} \\ \text{A2=PT}, & \text{if } y(\theta) \in \{\text{A2=PT}, \text{A2=PT}(\theta)\} \\ y(\theta), & \text{otherwise} \end{cases} \\
Y'(\theta)[y, z(\theta)] &\equiv \text{the set of path-equivalents } y'(\theta) \text{ that are possible for each field that follows the paths } y \text{ and } z(\theta)
\end{aligned}
\tag{3.39}
$$

For the remainder of this section, we will discuss $y'(\theta)$ in place of $y(\theta)$. After substituting $E_{\text{PT}(\theta)}$, the modified static model gives the following probabilities:

$$P[y'(\theta)] \equiv \text{the probability that a field follows path-equivalent } y'(\theta)$$

$$P[\text{PT=PD}] = (1 - E_{\text{PD}})(1 - E_{\text{PT}(\theta)}) \tag{3.40a}$$

$$P[\text{A1=PD}] = (1 - E_{\text{PD}})E_{\text{PT}(\theta)}(1 - E_{\text{A1}}) \tag{3.40b}$$

$$P[\text{A1=PT}] = E_{\text{PD}}(1 - E_{\text{PT}(\theta)})(1 - E_{\text{A1}}) \tag{3.40c}$$

$$P[\text{A2=PD}] = (1 - E_{\text{PD}})E_{\text{PT}(\theta)}E_{\text{A1}}(1 - E_{A2}) \tag{3.40d}$$

$$P[\text{A2=PT}] = E_{\text{PD}}(1 - E_{\text{PT}(\theta)})E_{\text{A1}}(1 - E_{\text{A2}}) \tag{3.40e}$$

$$P[\text{A2=A1}] = E_{\text{PD}}E_{\text{PT}(\theta)}(1 - E_{\text{A1}})(1 - E_{\text{A2}}) \tag{3.40f}$$

$$P[\text{INC}] = E_{\text{PD}}E_{\text{PT}(\theta)}(E_{\text{A1}} + E_{\text{A2}}) + E_{\text{A1}}E_{\text{A2}}(E_{\text{PD}} + E_{\text{PT}(\theta)}) - 3E_{\text{PD}}E_{\text{PT}(\theta)}E_{\text{A1}}E_{\text{A2}} \tag{3.40g}$$

$$\sum_{y'(\theta) \in Y} P[y'(\theta)] = 1 \tag{3.40h}$$

The field volumes are defined similarly:

$$f[y'(\theta)] \equiv \text{the number of fields that follow path-equivalent } y'(\theta)$$
$$F[y'(\theta)] \equiv \text{the rate at which fields follow path-equivalent } y'(\theta)$$
$$F[y'(\theta)] = \frac{f[y'(\theta)]}{f_{\mathrm{T}}} \tag{3.41}$$

Because the path-equivalent $y'(\theta)$ is indeterminate in some cases, we define the minimum and maximum volumes and rates as follows:

$$YZ(\theta)_1[y'(\theta)] \equiv \text{the subset of } YZ(\theta)_1 \text{ for which the predicted determinate outcome is } y'(\theta)$$
$$YZ(\theta)_1[y'(\theta)] = \{(y, z(\theta)) \in YZ(\theta)_1 : Y'(\theta)[y, z(\theta)] = \{y'(\theta)\}\} \tag{3.42}$$
$$YZ(\theta)_2[y'(\theta)] \equiv \text{the subset of } YZ(\theta)_2 \text{ for which the predicted indeterminate outcomes include } y'(\theta)$$
$$YZ(\theta)_2[y'(\theta)] = \{(y, z(\theta)) \in YZ(\theta)_2 : y'(\theta) \in Y(\theta)[y, z(\theta)]\} \tag{3.43}$$
$$\min f[y'(\theta)] = \sum_{(y,z(\theta)) \in YZ(\theta)_1[y'(\theta)]} f[y, z(\theta)] \tag{3.44}$$
$$\max f[y'(\theta)] = \min f[y'(\theta)] + \sum_{(y,z(\theta)) \in YZ(\theta)_2[y'(\theta)]} f[y, z(\theta)] \tag{3.45}$$
$$\min F[y'(\theta)] = \frac{\min f[y'(\theta)]}{f_{\mathrm{T}}} \tag{3.46}$$
$$\max F[y'(\theta)] = \frac{\max f[y'(\theta)]}{f_{\mathrm{T}}} \tag{3.47}$$

The following are also true:

$$\sum_{y'(\theta) \in Y} \min F[y'(\theta)] + \sum_{(y,z(\theta)) \in YZ(\theta)_2} F[y, z(\theta)] = 1 \tag{3.48}$$
$$\sum_{y'(\theta) \in Y} \min F[y'(\theta)] + F[\text{PT=PD}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4}] + F[\text{A1=PT}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4}] = 1 \tag{3.49}$$

We wish to compute a lower and upper bound on the Projected Truth error rate. In order to do so, we must find the following:

$$\min E2_{\mathrm{T}(\theta)} = \min \max_{y'(\theta) \in Y} |F[y'(\theta)] - P[y'(\theta)]| \tag{3.50}$$
$$\max E2_{\mathrm{T}(\theta)} = \max \max_{y'(\theta) \in Y} |F[y'(\theta)] - P[y'(\theta)]| \tag{3.51}$$

As indicated above, there are two sets of fields that must be distributed among the path-equivalents $y'(\theta)$ in such a way as to minimize or maximize the Projected Truth error rate. To see where the fields may be allocated, we examine the following:

$$Y'(\theta)[\text{PT=PD}, \text{R}_{\text{ACC}}(\theta)\text{=Other4}] = \{\text{A1=PD}, \text{A1=PT}, \text{A2=PD}, \text{A2=PT}, \text{A2=A1}, \text{INC}\} \quad (3.52)$$

$$Y'(\theta)[\text{A1=PT}, \text{R}_{\text{ACC}}(\theta)\text{=Other4}] = \{\text{A2=PD}, \text{A2=PT}, \text{A2=A1}, \text{INC}\} \quad (3.53)$$

$$Y'(\theta)[\text{A1=PT}, \text{R}_{\text{ACC}}(\theta)\text{=Other4}] \subset Y'(\theta)[\text{PT=PD}, \text{R}_{\text{ACC}}(\theta)\text{=Other4}] \quad (3.54)$$

In each case, we will first distribute the fields in the subset $Y'(\theta)[\text{A1=PT}, \text{R}_{\text{ACC}}(\theta)\text{=Other4}]$ optimally, followed by those in the superset $Y'(\theta)[\text{PT=PD}, \text{R}_{\text{ACC}}(\theta)\text{=Other4}]$. For the minimum case, we define the function MIN-PROJECTED-TRUTH-ERROR-RATE-2, along with its helper procedure MIN-DISTRIBUTE, as follows:

```
MIN-DISTRIBUTE(A, S, r)
 1  for each y'(θ) ∈ S
 2       do APPEND(A, min F[y'(θ)] − P[y'(θ)])
 3  SORT(A)
 4  i ← 1
 5  while i ≤ length[A] and r > 0
 6       do if i < length[A]
 7            then s ← MIN(A[i + 1] − A[i], r/i)
 8            else  s ← r/i
 9          for j ← 1 to i
10               do A[j] ← A[j] + s
11          r ← r − is
12          i ← i + 1
```

```
MIN-PROJECTED-TRUTH-ERROR-RATE-2()
1  A ← ⟨⟩
2  MIN-DISTRIBUTE(A, Y'(θ)[A1=PT, R_ACC(θ)=Other4], F[A1=PT, R_ACC(θ)=Other4])
3  MIN-DISTRIBUTE(
       A,
       Y'(θ)[PT=PD, R_ACC(θ)=Other4] − Y'(θ)[A1=PT, R_ACC(θ)=Other4],
       F[PT=PD, R_ACC(θ)=Other4]
   )
4  APPEND(A, F[PT=PD] − P[PT=PD])
5  return MAX-ABS(A)
```

For the maximum case, we define the function MAX-PROJECTED-TRUTH-ERROR-RATE-2, along with its helper procedure MAX-DISTRIBUTE, as follows:

```
MAX-DISTRIBUTE(A, S, r)
1  for each y'(θ) ∈ S
2       do APPEND(A, min F[y'(θ)] − P[y'(θ)])
3  i ← INDEX-OF-MAX(A)
4  A[i] ← A[i] + r

MAX-PROJECTED-TRUTH-ERROR-RATE-2()
1  A ← ⟨⟩
2  MAX-DISTRIBUTE(A, Y'(θ)[A1=PT, R_ACC(θ)=Other4], F[A1=PT, R_ACC(θ)=Other4])
3  MAX-DISTRIBUTE(
       A,
       Y'(θ)[PT=PD, R_ACC(θ)=Other4] − Y'(θ)[A1=PT, R_ACC(θ)=Other4],
       F[PT=PD, R_ACC(θ)=Other4]
   )
4  APPEND(A, F[PT=PD] − P[PT=PD])
5  return MAX-ABS(A)
```

Thus, we have the following definitions for the minimum and maximum Projected Truth error rate under Method 2:

$$\min E2_{\mathrm{T}(\theta)} = \text{MIN-PROJECTED-TRUTH-ERROR-RATE-2}() \tag{3.55}$$

$$\max E2_{\mathrm{T}(\theta)} = \text{MAX-PROJECTED-TRUTH-ERROR-RATE-2}() \tag{3.56}$$

3.6 Projected Manual Processing Rate

For each field, given the paths y, $z(\theta)$, and $y(\theta)$, we can determine exactly the incremental contribution to the projected manual processing volume.

We define the following functions in support of these calculations:

$$\varepsilon f_{\mathrm{K}(\theta)}[y, z(\theta), y(\theta)] \equiv \text{the incremental contribution to the Projected Keyer field count for each field that follows the paths } y, z(\theta), \text{ and } y(\theta)$$

$$\varepsilon f_{\mathrm{K}(\theta)}[y, z(\theta), y(\theta)] = \begin{cases} 1, & \text{if } z(\theta) = \mathrm{REJ}(\theta) \\ 0, & \text{otherwise} \end{cases} \tag{3.57}$$

$$\varepsilon f_{\mathrm{A1}(\theta)}[y, z(\theta), y(\theta)] \equiv \text{the incremental contribution to the Projected Analyst 1 field count for each field that follows the paths } y, z(\theta), \text{ and } y(\theta)$$

$$\varepsilon f_{\mathrm{A1}(\theta)}[y, z(\theta), y(\theta)] = \begin{cases} 1, & \text{if } y(\theta) \notin \{\mathrm{PT{=}PD}, \mathrm{PT}(\theta){=}\mathrm{PD}\} \\ 0, & \text{otherwise} \end{cases} \tag{3.58}$$

$$\varepsilon f_{\mathrm{A2}(\theta)}[y, z(\theta), y(\theta)] \equiv \text{the incremental contribution to the Projected Analyst 2 field count for each field that follows the paths } y, z(\theta), \text{ and } y(\theta)$$

$$\varepsilon f_{\mathrm{A2}(\theta)}[y, z(\theta), y(\theta)] = \begin{cases} 1, & \text{if } y(\theta) \in \{\mathrm{A2{=}PD}, \mathrm{A2{=}PT}, \mathrm{A2{=}PT}(\theta), \mathrm{A2{=}A1}, \mathrm{INC}\} \\ 0, & \text{otherwise} \end{cases} \tag{3.59}$$

Further, we can define the following determinate, minimum, and maximum incremental contributions to each projected data set, given the paths y

and $z(\theta)$:

$$\varepsilon f_{\mathrm{K}(\theta)}[y, z(\theta)] = \begin{cases} 1, & \text{if } z(\theta) = \mathrm{REJ}(\theta) \\ 0, & \text{otherwise} \end{cases} \tag{3.60}$$

$$\varepsilon f_{\mathrm{A1}(\theta)}[y, z(\theta)] = \begin{cases} 1, & \text{if } (y, z(\theta)) \in YZ(\theta)_1 \text{ and } Y(\theta)[y, z(\theta)] \notin \{\{\text{PT=PD}\}, \\ & \qquad \{\text{PT}(\theta)\text{=PD}\}\} \\ 1, & \text{if } (y, z(\theta)) \in YZ(\theta)_2 \\ 0, & \text{otherwise} \end{cases} \tag{3.61}$$

$$\min \varepsilon f_{\mathrm{A2}(\theta)}[y, z(\theta)] = \begin{cases} 1, & \text{if } Y(\theta)[y, z(\theta)] \in \{\{\text{A2=PD}\}, \{\text{A2=PT}\}, \{\text{A2=PT}(\theta)\}, \{\text{A2=A1}\}, \\ & \qquad \{\text{INC}\}\} \\ 0, & \text{if } (y, z(\theta)) = (\text{PT=PD}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4}) \\ 1, & \text{if } (y, z(\theta)) = (\text{A1=PT}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4}) \\ 0, & \text{otherwise} \end{cases} \tag{3.62}$$

$$\max \varepsilon f_{\mathrm{A2}(\theta)}[y, z(\theta)] = \begin{cases} 1, & \text{if } Y(\theta)[y, z(\theta)] \in \{\{\text{A2=PD}\}, \{\text{A2=PT}\}, \{\text{A2=PT}(\theta)\}, \{\text{A2=A1}\}, \\ & \qquad \{\text{INC}\}\} \\ 1, & \text{if } (y, z(\theta)) = (\text{PT=PD}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4}) \\ 1, & \text{if } (y, z(\theta)) = (\text{A1=PT}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4}) \\ 0, & \text{otherwise} \end{cases} \tag{3.63}$$

To determine the total contribution to the each projected data set, we multiply the incremental contribution by the number of fields that follow the paths y and $z(\theta)$:

$$f_{\mathrm{K}(\theta)}[y, z(\theta)] = \varepsilon f_{\mathrm{K}(\theta)}[y, z(\theta)] f[y, z(\theta)] \tag{3.64}$$

$$f_{\mathrm{A1}(\theta)}[y, z(\theta)] = \varepsilon f_{\mathrm{A1}(\theta)}[y, z(\theta)] f[y, z(\theta)] \tag{3.65}$$

$$\min f_{\mathrm{A2}(\theta)}[y, z(\theta)] = \min \varepsilon f_{\mathrm{A2}(\theta)}[y, z(\theta)] f[y, z(\theta)] \tag{3.66}$$

$$\max f_{\mathrm{A2}(\theta)}[y, z(\theta)] = \max \varepsilon f_{\mathrm{A2}(\theta)}[y, z(\theta)] f[y, z(\theta)] \tag{3.67}$$

Then, we obtain the total field count of each projected data set as follows:

$$f_{\mathrm{K}(\theta)} = \sum_{y \in Y} \sum_{z(\theta) \in Z(\theta)} f_{\mathrm{K}(\theta)}[y, z(\theta)] \tag{3.68}$$

$$f_{\mathrm{A1}(\theta)} = \sum_{y \in Y} \sum_{z(\theta) \in Z(\theta)} f_{\mathrm{A1}(\theta)}[y, z(\theta)] \tag{3.69}$$

$$\min f_{\mathrm{A2}(\theta)} = \sum_{y \in Y} \sum_{z(\theta) \in Z(\theta)} \min f_{\mathrm{A2}(\theta)}[y, z(\theta)] \tag{3.70}$$

$$\max f_{\mathrm{A2}(\theta)} = \sum_{y \in Y} \sum_{z(\theta) \in Z(\theta)} \max f_{\mathrm{A2}(\theta)}[y, z(\theta)] \tag{3.71}$$

$$\max f_{\mathrm{A2}(\theta)} = \min f_{\mathrm{A2}(\theta)} + f[\mathrm{PT{=}PD}, \mathrm{R_{ACC}}(\theta){=}\mathrm{Other4}] \tag{3.72}$$

Finally, we have definitions for the minimum and maximum projected manual processing volume and rate:

$$\min m_{\mathrm{T}(\theta)} = f_{\mathrm{K}(\theta)} + f_{\mathrm{A1}(\theta)} + \min f_{\mathrm{A2}(\theta)} \tag{3.73}$$

$$\max m_{\mathrm{T}(\theta)} = f_{\mathrm{K}(\theta)} + f_{\mathrm{A1}(\theta)} + \max f_{\mathrm{A2}(\theta)} \tag{3.74}$$

$$\max m_{\mathrm{T}(\theta)} = \min m_{\mathrm{T}(\theta)} + f[\mathrm{PT{=}PD}, \mathrm{R_{ACC}}(\theta){=}\mathrm{Other4}] \tag{3.75}$$

$$\min M_{\mathrm{T}(\theta)} = \frac{\min m_{\mathrm{T}(\theta)}}{f_{\mathrm{T}}} \tag{3.76}$$

$$\max M_{\mathrm{T}(\theta)} = \frac{\max m_{\mathrm{T}(\theta)}}{f_{\mathrm{T}}} \tag{3.77}$$

$$\max M_{\mathrm{T}(\theta)} = \min M_{\mathrm{T}(\theta)} + F[\mathrm{PT{=}PD}, \mathrm{R_{ACC}}(\theta){=}\mathrm{Other4}] \tag{3.78}$$

3.7 Practical Tuning Application

If we consider a PDQ instance that has just begun operations, we first want to know when we have processed a large enough sample to make any tuning decisions. Paxton [10] has written some guidelines that pertain directly to assessing the Production Data Capture system, but we can apply similar principles in assessing PDQ's own data quality. To begin, we define some

new terms:

$E_{\text{PD}}^{ref} \equiv$ the reference, or target, error rate of the Production Data set

$E_{\text{T}}^{ref} \equiv$ the reference error rate of the Truth data set

$D \equiv$ the desired ratio of the reference Production Data error rate to the reference Truth error rate

$$D = \frac{E_{\text{PD}}^{ref}}{E_{\text{T}}^{ref}} \tag{3.79}$$

$\sigma_{\text{T}}^{ref} \equiv$ the standard error associated with the reference Truth error rate

$d \equiv$ the desired ratio of the reference Truth error rate to the associated standard error

$$d = \frac{E_{\text{T}}^{ref}}{\sigma_{\text{T}}^{ref}} \tag{3.80}$$

Assuming that the PDQ sample is very small compared to the population, we can use a simplified estimate for standard error:

$\sigma_x \equiv$ the estimated standard error associated with the error rate of data set x

$$\sigma_x = \sqrt{\frac{E_x(1 - E_x)}{f_x}} \tag{3.81}$$

Given the above equations, we can solve for the reference sample size f_{T}^{ref} in terms of parameters E_{PD}^{ref}, D, and d:

$$f_{\text{T}}^{ref} = \frac{Dd^2}{E_{\text{PD}}^{ref}} - d^2 \tag{3.82}$$

Once we have reached this reference sample size within PDQ, then we are ready to perform some analysis of the Truth error rate. First, we analyze the current performance of PDQ by applying a confidence interval to the estimated Truth error rate [5]:

$c \equiv$ the desired confidence, expressed as a fraction or percentage, that the estimated Truth error rate is at or below the reference Truth error rate

$z_c \equiv$ the value given by the probit function for the probability c

$$z_c = \Phi^{-1}(c) \tag{3.83}$$

We can then test whether the following is true:

$$E_{\mathrm{T}} + z_c \sigma_{\mathrm{T}} \leq E_{\mathrm{T}}^{ref} \tag{3.84}$$

For this application, we will assume the case that the test passes. We can then use one of the Projected Truth error rate estimates. We can find the smallest value of θ that satisfies the following:

$$E_{\mathrm{T}(\theta)} + z_c \sigma_{\mathrm{T}(\theta)} \leq E_{\mathrm{T}}^{ref} \tag{3.85}$$

Once we find the desired operating point, we can reconfigure the Independent Data Capture system's confidence threshold. The corresponding projected manual processing rate allows us to plan future capacity. This approach is similar to evaluating the error and reject tradeoff in the Recognizer, but unlike the latter, the Projector model accounts for the performance characteristics of the complete PDQ system.

Chapter 4

Evaluation Methods

4.1 Experimental Data

For the 2010 Census, an instance of PDQ processed a sample of about 865 thousand paper forms (rather, images thereof) in order to estimate the data capture quality of the roughly 164 million forms processed by the DRIS contractor [7]. The forms in the sample represent nearly 50 form types, which were used for various DRIS operations and targeted at different population segments. The form types vary in question language, background color, and expected marking instrument, among other factors that have fundamental impacts on data capture quality. PDQ processed and analyzed both write-in and check-box fields on these forms; write-in fields were further classified as alphabetic, numeric, and alphanumeric, depending on the allowed character sets.

Given these distinctions, I have selected a subset of the PDQ sample consisting of numeric write-in fields on D-1(E) forms, which were the main form type used by Census enumerators for the Nonresponse Followup (NRFU) operation. The 278,639 D-1(E) forms constitute the largest form count of any form type in the PDQ sample, and numeric fields are the most prevalent subtype of write-in fields. Certain special-purpose features, particularly manual identification of a small number of ambiguous or erased fields, were added to PDQ for the 2010 Census. These features introduced additional process flow considerations that have not been represented in Section 2.1, so I have excluded any fields that were impacted by these features. Thus the 5,355,398 D-1(E) numeric fields examined in this study comprise a large stratum of work units that we expect to have reasonably

consistent behavior in the data capture process.

Due to strict security protocols surrounding the Census data, I have retrieved only aggregate statistical data, sufficient to provide the necessary inputs to the static model and Projector model. My selected sample spans PDQ processing dates from April 2, 2010, through September 30, 2010, and I have further stratified the data to allow evaluation of subsamples from specific months and weeks.

4.2 Analytical Approach

4.2.1 Overview

For 2010 Census operations, PDQ's Independent Data Capture system was configured with the confidence threshold $\theta_0 = 80$ across all fields. This was a relatively conservative decision that assured high data quality in the majority of cases, and there was no practical impetus for revisiting the position during that PDQ instance's operational lifetime.

Within the database query used to gather statistics for the selected sample, I have implemented the logical components of the Projector up to, but excluding, the Path Predictor. As a result, each subset of fields in the sample is identifiable by the original terminal path y, the hypothetical confidence threshold $\theta \in [-1, 80]$, and the implied intermediate path $z(\theta)$. Subsequently, I have implemented the remaining functions and calculations of the static model and Projector model via formulas in a Microsoft Excel workbook.

4.2.2 Assumptions and Limitations

This study depends on a number of practical assumptions. First, there is no reliable way to measure directly the error rates of Analyst 1 (E_{A1}) and Analyst 2 (E_{A2}) using the PDQ Truth. The Analyst 1 step itself is typically responsible for determining the Truth for more than 97% of the fields it encounters [7]. Analyst 2 designates the Truth value for *all* the fields it encounters. However, as implemented for the 2010 Census, PDQ drew from

a single pool of individuals for the Keyer, Analyst 1, and Analyst 2 steps, and the system has ensured that any given person filled at most one of these roles on a given form. Therefore, in all calculations, I substitute E_{K} — which is more reliably measured using the Truth — for both E_{A1} and E_{A2}.

As is customary when training or tuning a pattern classifier, we assume that the outputs of the Projector model, based on *past* inputs, are suitable for predicting *future* outcomes [6]. Because both DRIS and PDQ are complex, dynamic systems, this generalization does not hold perfectly. In the following chapter, we shall see an example of this issue, as well as a proposed method of dealing with it.

There is currently no operational instance of PDQ with which to test the predictions made by the Projector model. Thus I rely solely on analysis of historical data to draw conclusions about the usefulness of the model.

The data quality measurements presented here have been computed expressly for the purpose of understanding PDQ performance on a particular selected sample. Any references to DRIS data quality, specifically expressed as E_{PD}, do not reflect the official scores provided by PDQ for the 2010 Census, and should not be interpreted as such.

Chapter 5

Results

5.1 Static Model

Table 5.1 shows the results obtained for the total sample via the static model. Note that the estimated Truth error rate $E_{\mathrm{T}} = 0.00357\%$, which is nearly 160 times as low as the Production Data error rate ($E_{\mathrm{PD}} = 0.56724\%$). Clearly, the Truth is more than precise enough for the purpose of assessing the Production Data Capture system's data quality. While more than 98% of the fields bypassed the Arbitrator ($F[\mathrm{PT{=}PD}] = 98.19259\%$), the measured manual processing rate $M_{\mathrm{T}} = 33.23876\%$, most of which is comprised of reject keying within the Independent Data Capture system. As we examine the results from the Projector model, we will look for potential opportunities to improve PDQ's efficiency, while maintaining sufficient overall data quality.

5.2 Projector Model

5.2.1 Projected Truth Error Rate: Method 1

First, we consider the Projected Truth error rates given by Method 1. Figure 5.1 shows the various components of the minimum estimate $\min E1_{\mathrm{T}(\theta)}$, as functions of the confidence threshold θ. (Note that this chart and the ones that follow show connected data points for the sake of visibility, but the independent variable θ is always an integer.) As defined explicitly for this method, the projected error rate increases as the confidence threshold decreases. There is a small, constant contribution from

Table 5.1: Results for total sample: Static model.

(a) Sample size.

Form Count	278,639
f_T	5,355,398

(b) Inputs.

E_{PD}	0.56724%
E_{PT}	1.25083%
E_{A1}	0.39523%
E_{A2}	0.39523%

(c) Probabilities, volumes, and Truth error rate.

y	$P[y]$	$F[y]$	$F[y] - P[y]$	$\lvert F[y] - P[y] \rvert$
PT=PD	98.18902%	98.19259%	0.00357%	0.00357%
A1=PD	1.23882%	1.23825%	-0.00057%	0.00057%
A1=PT	0.55793%	0.55559%	-0.00234%	0.00234%
A2=PD	0.00490%	0.00192%	-0.00297%	0.00297%
A2=PT	0.00221%	0.00099%	-0.00122%	0.00122%
A2=A1	0.00704%	0.01053%	0.00349%	0.00349%
INC	0.00008%	0.00013%	0.00005%	0.00005%
				E_{T} = 0.00357%
				σ_{T} = 0.00026%
				$E_{\text{T}} + 1.645\sigma_{\text{T}}$ = 0.00399%

(d) Manual processing rate.

F_{K}	31.41778%
F_{A1}	1.80741%
F_{A2}	0.01358%
M_{T}	33.23876%

$E_{\mathrm{T}(\theta)}$, the Truth error rate given by the static model. The largest error component here is $\Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=PD}]$; that is, most of the Projected Truth errors in this estimate are incurred by newly-accepted fields whose values match incorrect Production Data values. In contrast, the remaining error components $\Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A1}]$, $\Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=A2}]$, and $\min \Delta E_{\mathrm{T}(\theta)}[\mathrm{R_{ACC}}(\theta)\text{=Other4}]$ are negligible.

Figure 5.2 shows the components of the maximum estimate $\max E1_{\mathrm{T}(\theta)}$. While there is a noticable contribution from $F[\text{A1=PT}, \mathrm{R_{ACC}}(\theta)\text{=Other4}]$, by far the largest error component is $F[\text{PT=PD}, \mathrm{R_{ACC}}(\theta)\text{=Other4}]$. As the confidence threshold decreases, a substantial portion of fields follows the paths $(y, z(\theta)) = (\text{PT=PD}, \mathrm{R_{ACC}}(\theta)\text{=Other4})$, which yield an indeterminate projected terminal path $y(\theta)$. Assuming that all these fields incur additional Truth errors gives a maximum estimate that increases sharply in contrast to the minimum estimate.

In Figure 5.3 we see a clear comparison of these minimum and maximum estimates.

5.2.2 Projected Provisional Truth Error Rate

Figure 5.4 shows the components of the projected Provisional Truth error rate $E_{\mathrm{PT}(\theta)}$, as functions of the confidence threshold θ. Overall, the projected error rate increases as the confidence threshold decreases. At all points, the largest error component is $E_{\mathrm{R_{ACC}}(\theta)}$, which is contributed by accepted fields. The other error component, $E_{\mathrm{K}(\theta)}$, consists of projected Keyer errors. This component starts at less than one tenth of the total at $\theta = \theta_0 = 80$, and it decreases slowly as more fields become accepted.

5.2.3 Projected Truth Error Rate: Method 2

Next, we examine the Projected Truth error rates given by Method 2. Figure 5.5 shows the minimum and maximum estimates ($\min E2_{\mathrm{T}(\theta)}$ and $\max E2_{\mathrm{T}(\theta)}$, respectively), as functions of the confidence threshold θ. Consistent with the other projected error rates observed so far, these estimates increase as the confidence threshold decreases.

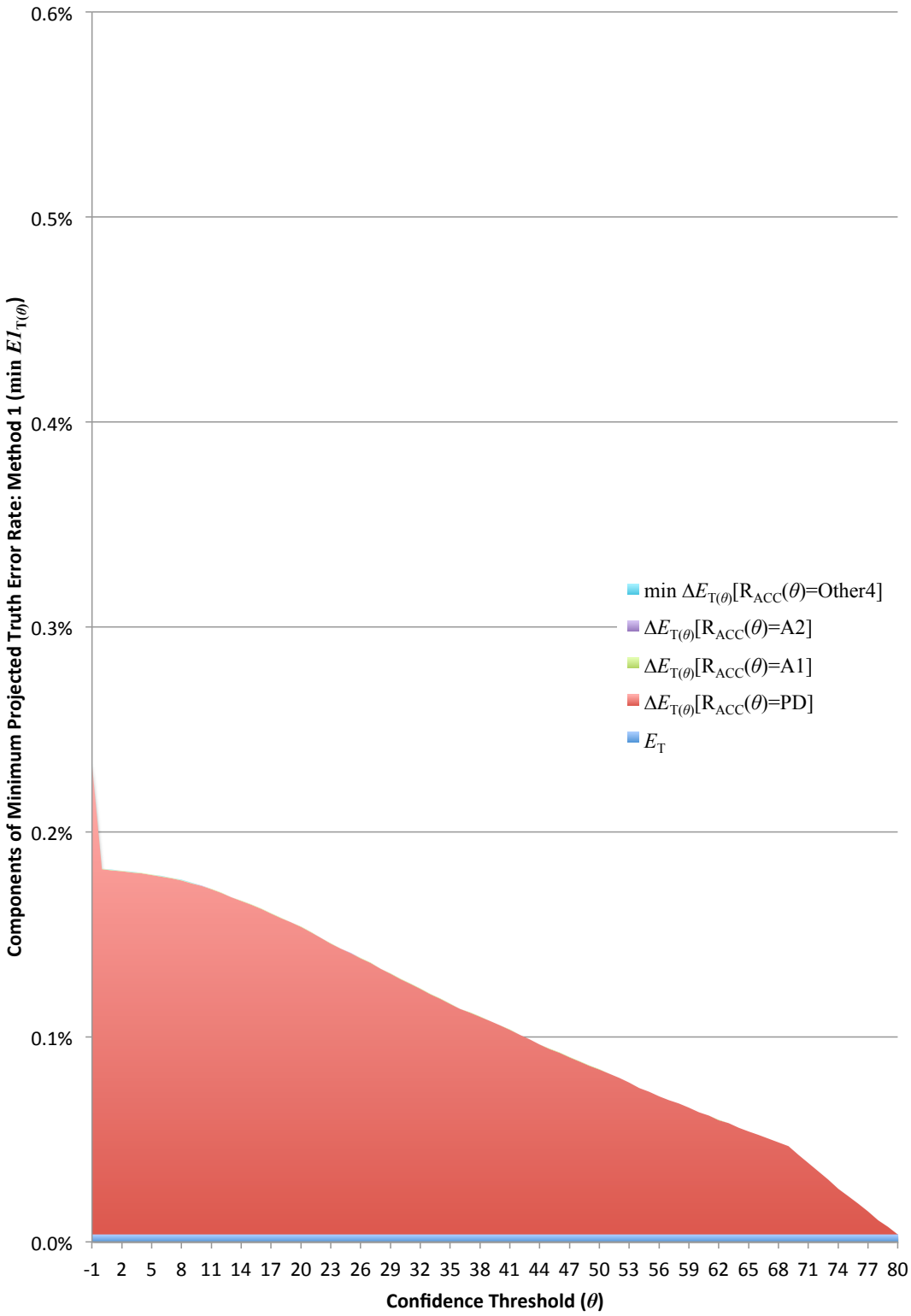


Figure 5.1: Results for total sample: Components of minimum Projected Truth error rate: Method 1 ($\min E1_{T(\theta)}$).

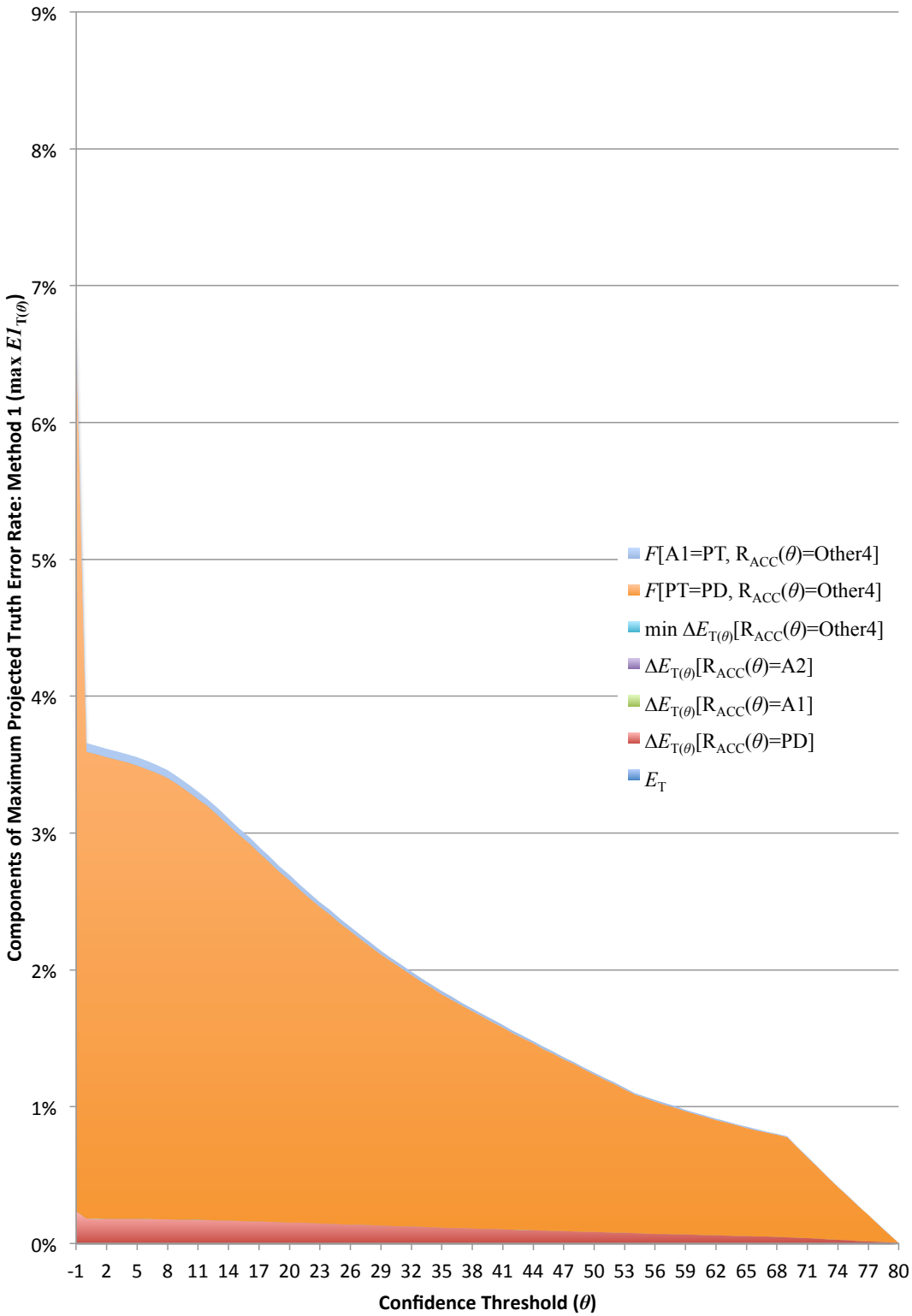


Figure 5.2: Results for total sample: Components of maximum Projected Truth error rate: Method 1 ($\max EI_{T(\theta)}$).

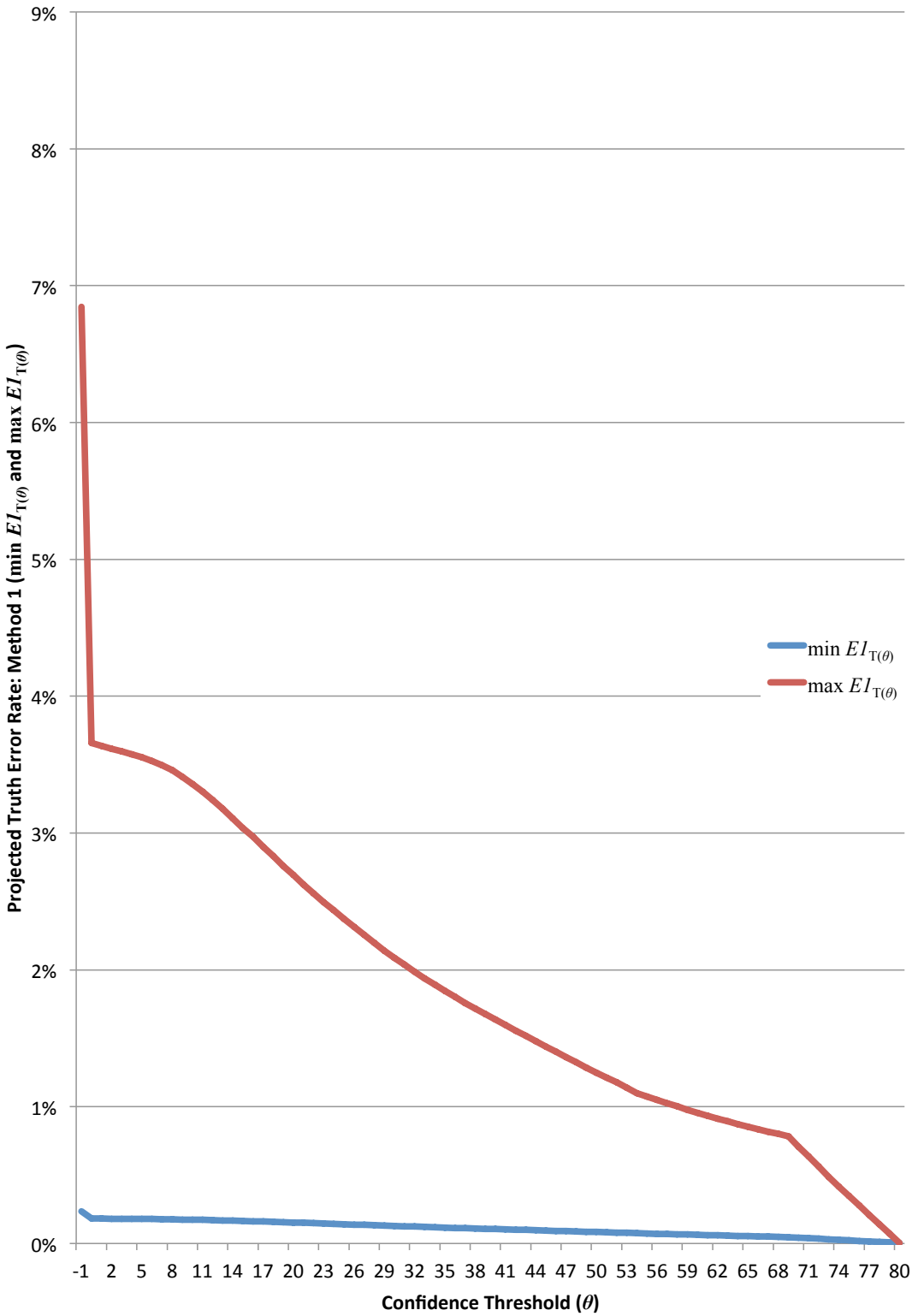


Figure 5.3: Results for total sample: Comparison of minimum and maximum Projected Truth error rates: Method 1 ($\min E1_{T(\theta)}$ and $\max E1_{T(\theta)}$).

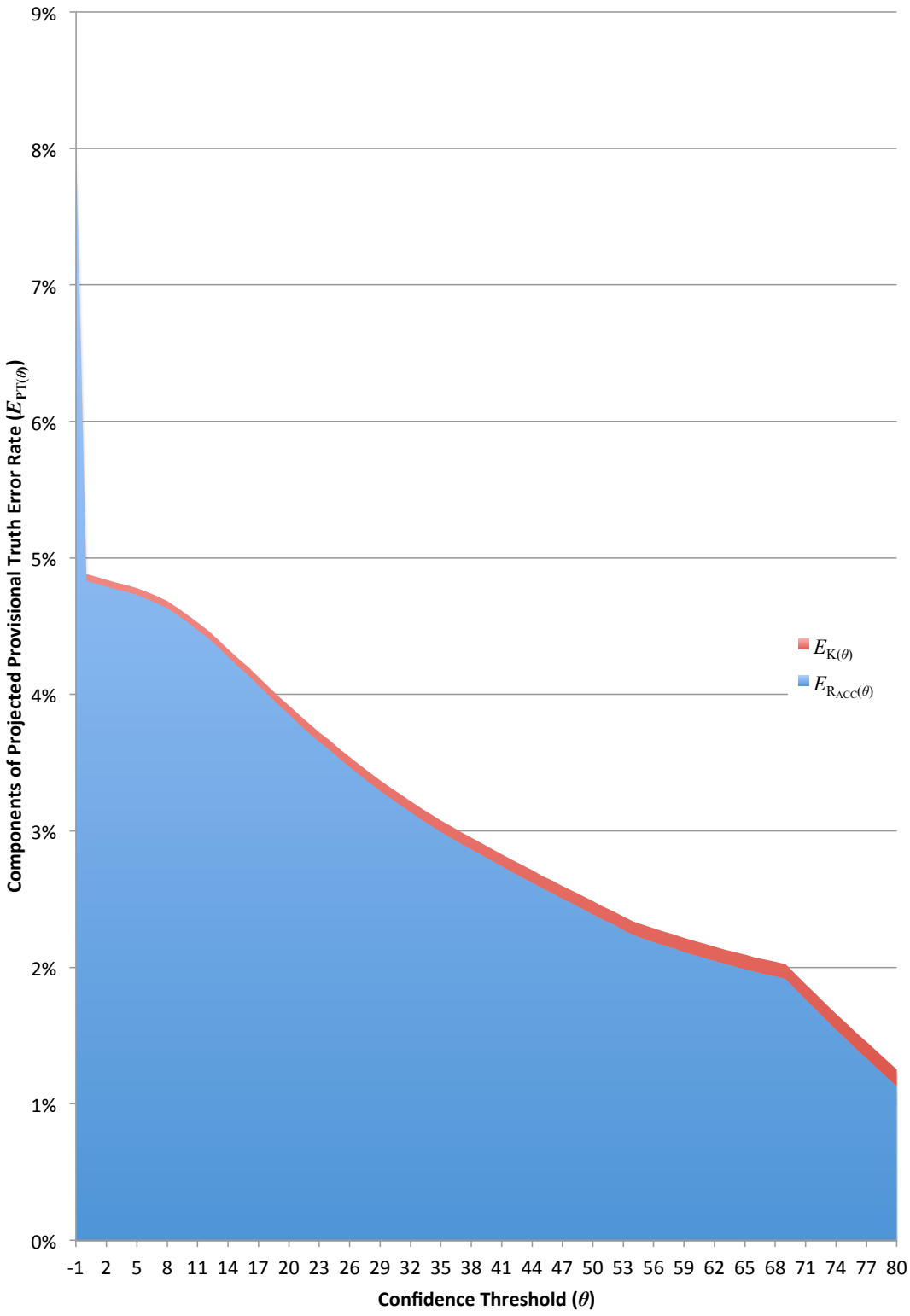


Figure 5.4: Results for total sample: Components of Projected Provisional Truth error rate ($E_{\mathrm{PT}(\theta)}$).

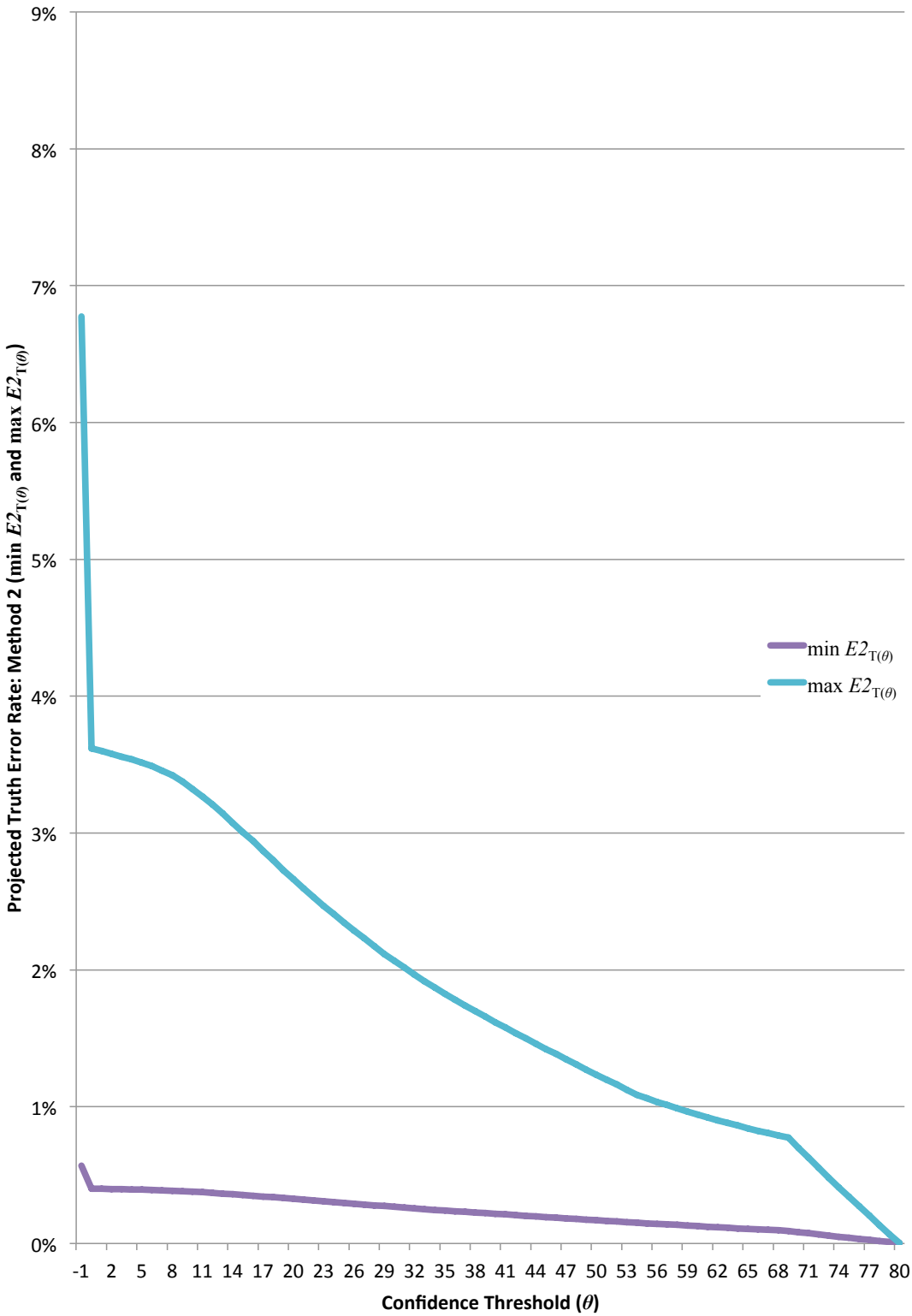


Figure 5.5: Results for total sample: Comparison of minimum and maximum Projected Truth error rates: Method 2 ($\min E2_{T(\theta)}$ and $\max E2_{T(\theta)}$).

Figure 5.6 compares the various projected error rate estimates. We see immediately that there is close agreement between the respective minimum and maximum Projected Truth error rates given by Methods 1 and 2. Note that $\min E1_{\mathrm{T}(\theta)}$ and $\min E2_{\mathrm{T}(\theta)}$ are within the same order of magnitude, and that $\max E1_{\mathrm{T}(\theta)}$ and $\max E2_{\mathrm{T}(\theta)}$ are very nearly equal. This evidence supports the validity of Method 1 and Method 2, and of the static model, upon which Method 2 is based directly.

Curiously, there is an almost constant difference between the Projected Provisional Truth error rate $E_{\mathrm{PT}(\theta)}$ and either of the maximum Projected Truth error rates $\max E1_{\mathrm{T}(\theta)}$ and $\max E2_{\mathrm{T}(\theta)}$. This relationship suggests that the maximum estimates reflect truly "worst-case" scenarios, in which the Truth error rate is completely dependent upon the Provisional Truth error rate. We can conclude that these maximum estimates are poorly suited for realistic evaluation of PDQ's performance.

Henceforth, I will use $\min E2_{\mathrm{T}(\theta)}$ as the preferred estimate for practical purposes.

5.2.4 Projected Manual Processing Rate

We turn now to the estimates of projected manual processing rate given by the Projector model. Figure 5.7 shows the various components of the minimum estimate $\min M_{\mathrm{T}(\theta)}$, as functions of the confidence threshold θ. Overall, the projected manual processing rate decreases with the confidence threshold. The largest manual processing component is the reject rate $F_{\mathrm{K}(\theta)}$ at $\theta = \theta_0 = 80$, while a greater proportion shifts toward Analyst 1 (F_{A1}) at lower thresholds. The remaining component, $\min F_{\mathrm{A2}}$, is negligible. Note that unlike the reject rate, the overall minimum projected manual processing rate has a non-zero lower bound ($\min M_{\mathrm{T}(-1)} = 8.16987\%$).

Figure 5.8 shows the components of the maximum projected manual processing rate $\max M_{\mathrm{T}(\theta)}$. The additional component $F[\text{PT=PD}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4}]$ contributes a significant portion of the overall manual processing as the confidence threshold decreases. In this estimate, all fields that follow the paths $(y, z(\theta)) = (\text{PT=PD}, \mathrm{R}_{\mathrm{ACC}}(\theta)\text{=Other4})$, which yield an indeterminate projected terminal path $y(\theta)$, are assumed to

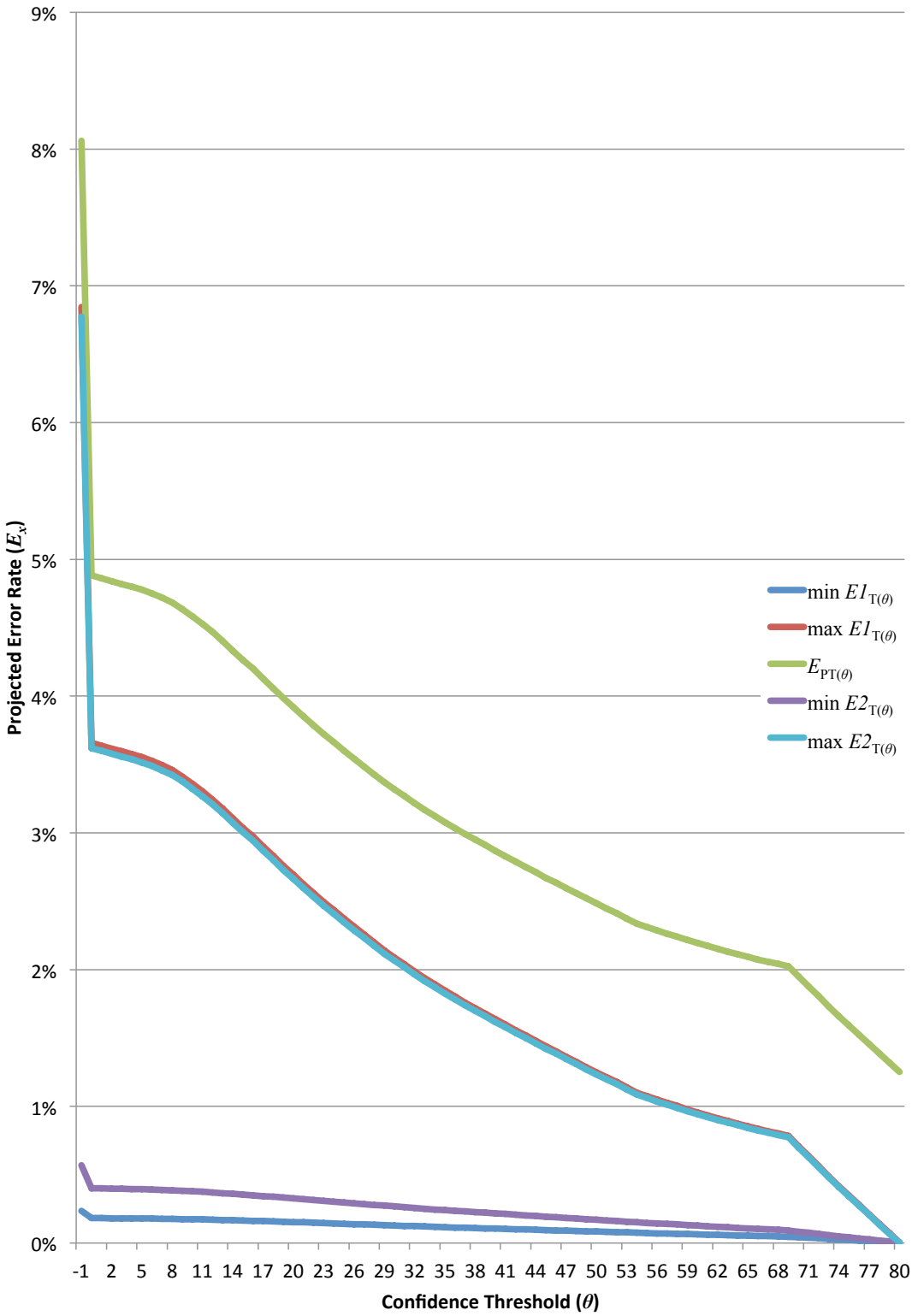


Figure 5.6: Results for total sample: Comparison of projected error rates (E_x).

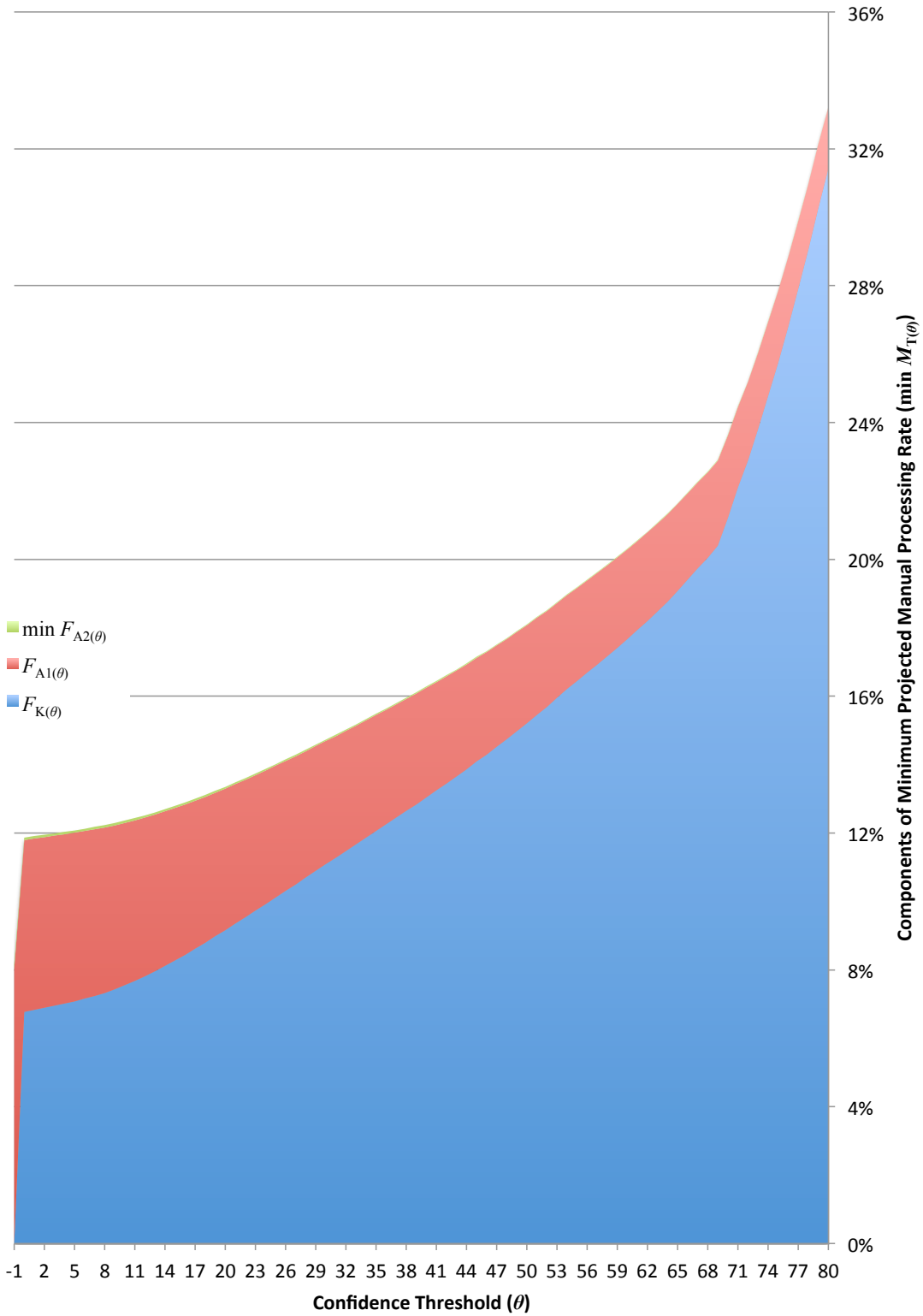


Figure 5.7: Results for total sample: Components of minimum projected manual processing rate ($\min M_{T(\theta)}$).

incur processing by Analyst 2.

Figure 5.9 shows a clear comparison of $\min M_{\mathrm{T}(\theta)}$ and $\max M_{\mathrm{T}(\theta)}$. In contrast to the relationships between $\min E1_{\mathrm{T}(\theta)}$ and $\max E1_{\mathrm{T}(\theta)}$ or between $\min E2_{\mathrm{T}(\theta)}$ and $\max E2_{\mathrm{T}(\theta)}$, the minimum and maximum projected manual processing rates are within the same order of magnitude at all θ values.

In further analysis and discussion, I will use $\max M_{\mathrm{T}(\theta)}$ as the preferred, conservative estimate for practical purposes.

5.3 Error and Manual Processing Tradeoff

In Figure 5.10, we see a classic error ($E_{\mathrm{PT}(\theta)}$) v. reject ($F_{\mathrm{K}(\theta)}$) curve for the Independent Data Capture system, showing a characteristic tradeoff between the two metrics.

Figure 5.11 shows an analogous relationship between Truth error ($\min E2_{\mathrm{T}(\theta)}$) and manual processing ($\max M_{\mathrm{T}(\theta)}$). The exception, as noted before, is that the manual processing rate reaches a non-zero minimum.

While the two visualizations serve equivalent roles for the Independent Data Capture system and for PDQ as a whole, there does not appear to be a simple means to relate one to the other. Figure 5.12 shows an attempt to compare the two curves by setting their respective vertical scales to similar proportions. This demonstrates that it is insufficient to rely on the performance characteristics of the Independent Data Capture system for tuning operating parameters such as the confidence threshold. To achieve successful outcomes, it is essential to understand the relationship between the overall performance characteristics of PDQ.

5.4 Practical Tuning Application

For the 2010 Census Production Data Capture system, write-in fields accepted by the OCR engine were required to have a maximum error rate of 1%, and write-in fields sent to reject keying were required to have a maximum error rate of 3%. There was also a design goal to keep the reject rate below 20%. Given these values, we can compute a weighted reference error

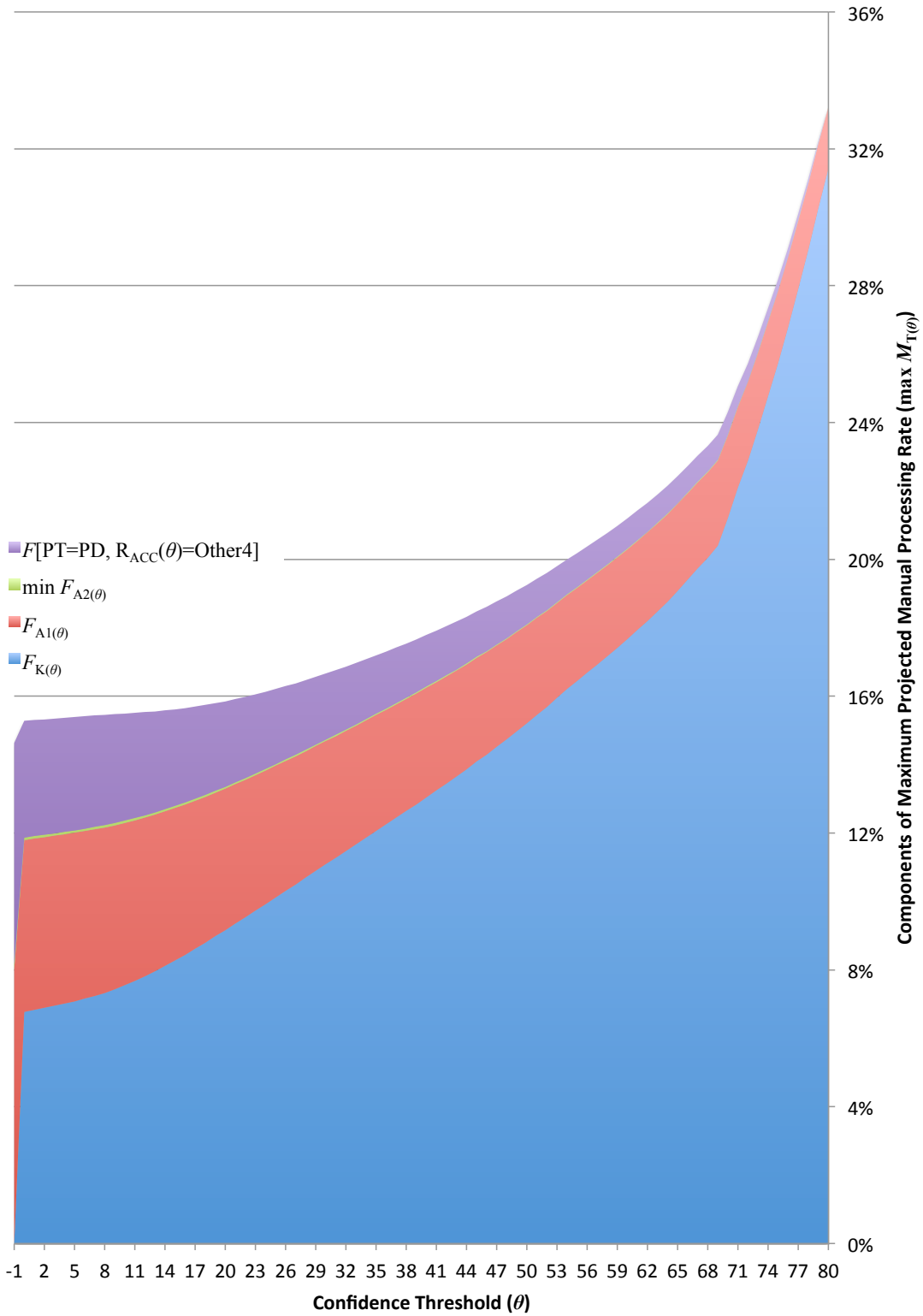


Figure 5.8: Results for total sample: Components of maximum projected manual processing rate ($\max M_{T(\theta)}$).

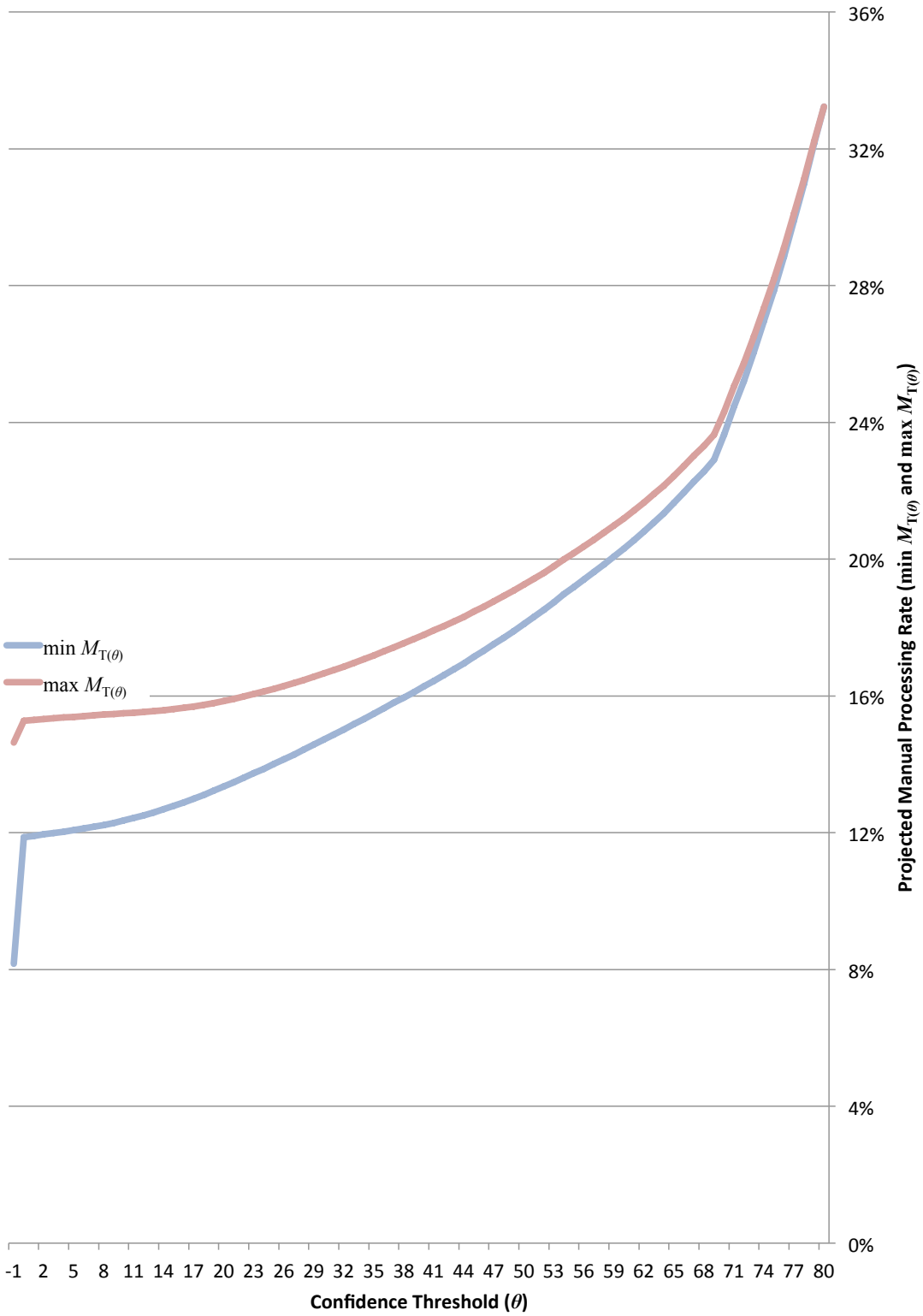


Figure 5.9: Results for total sample: Comparison of minimum and maximum projected manual processing rates ($\min M_{\mathrm{T}(\theta)}$ and $\max M_{\mathrm{T}(\theta)}$).

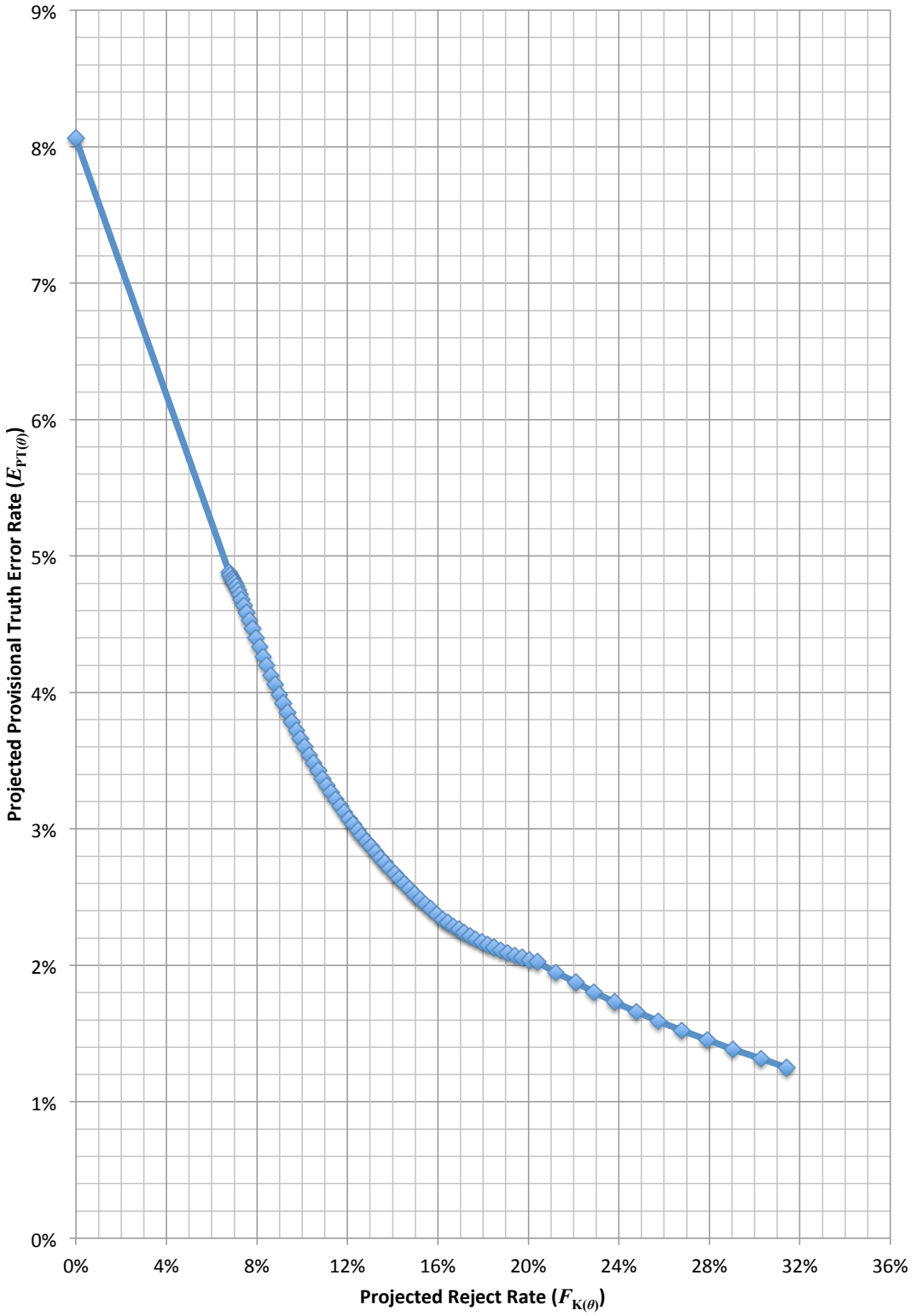


Figure 5.10: Results for total sample: Projected Provisional Truth error rate ($E_{\mathrm{PT}(\theta)}$) v. projected reject rate ($F_{\mathrm{K}(\theta)}$).

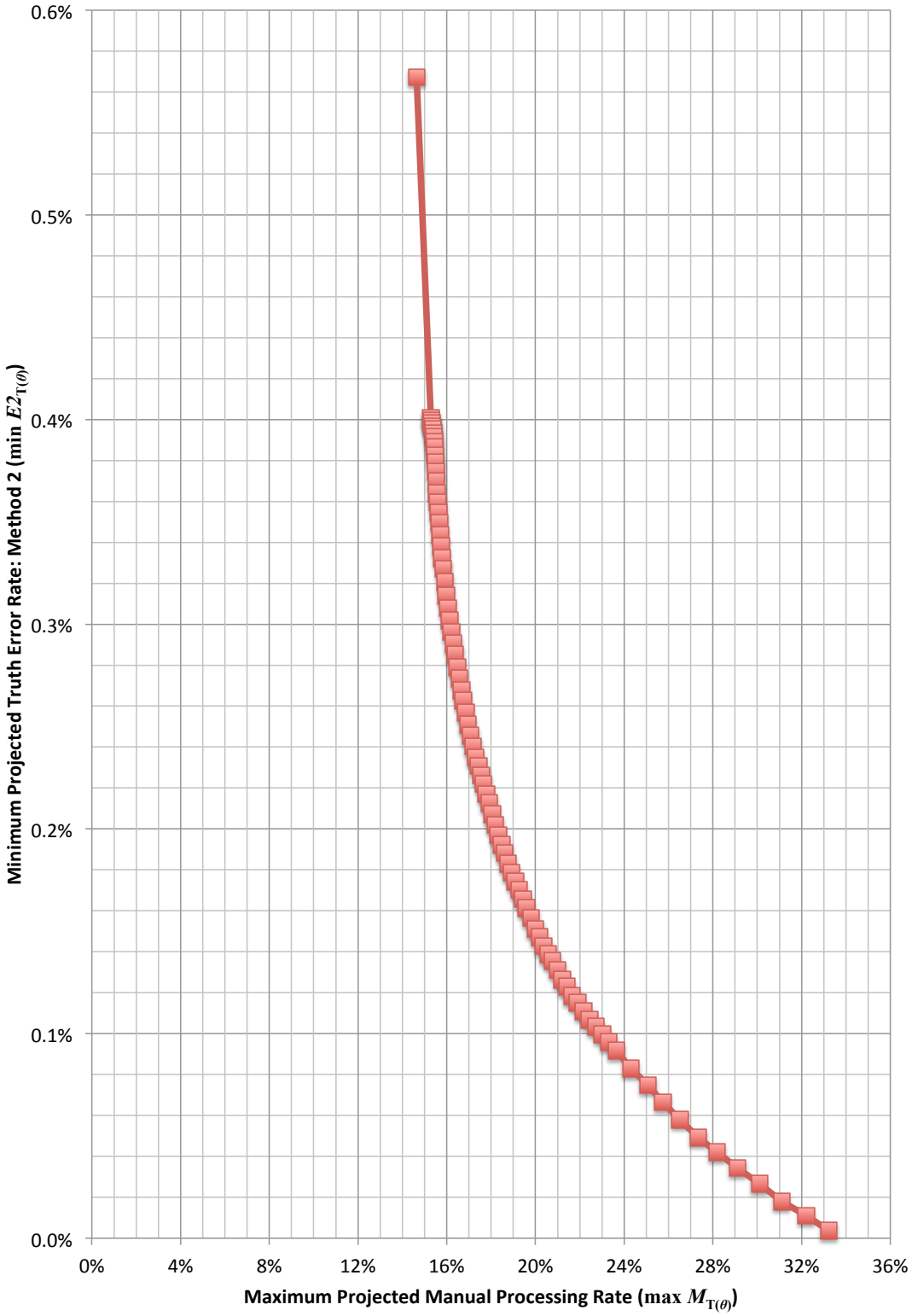


Figure 5.11: Results for total sample: Minimum Projected Truth error rate: Method 2 ($\min E2_{\mathrm{T}(\theta)}$) v. maximum projected manual processing rate ($\max M_{\mathrm{T}(\theta)}$).

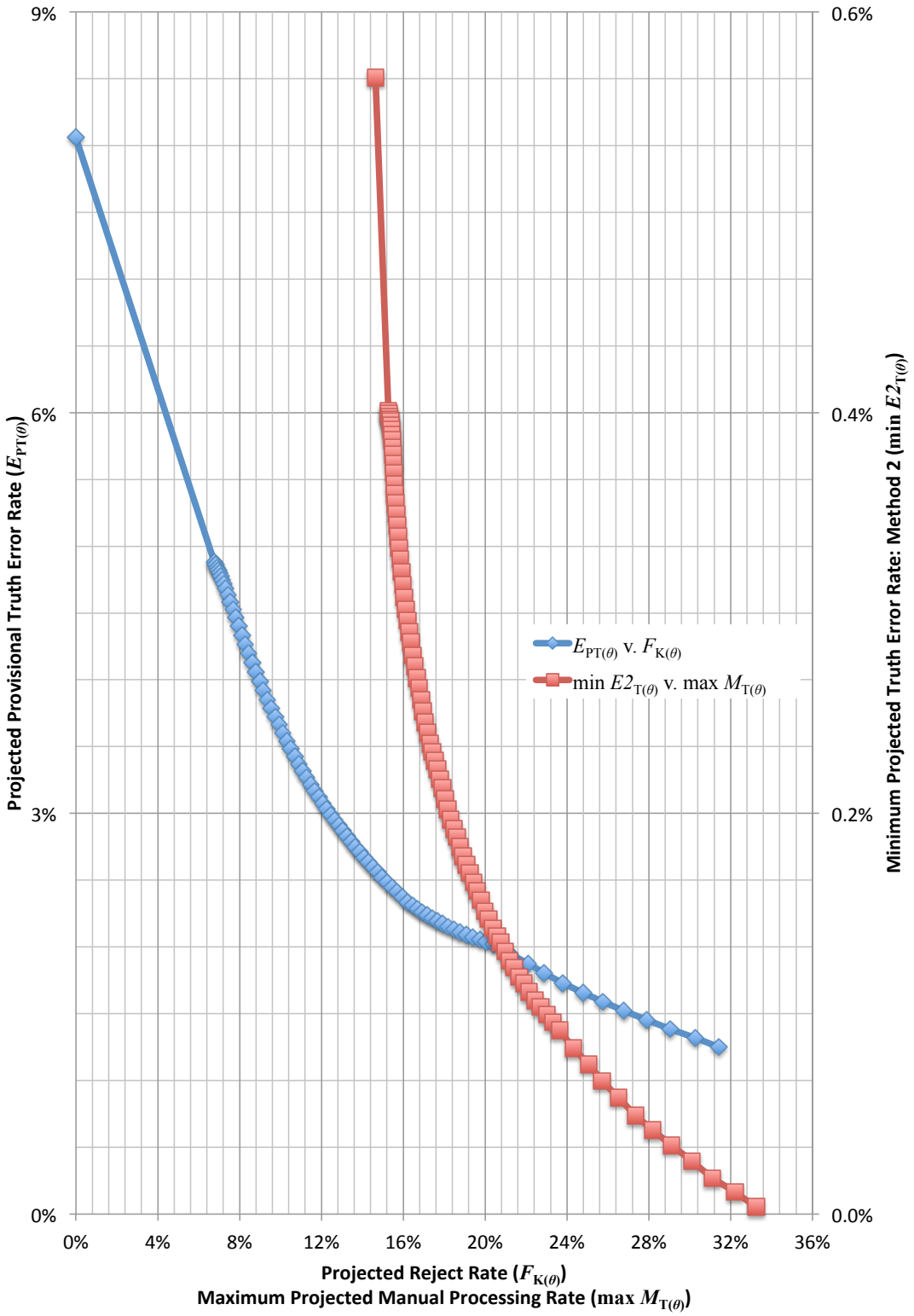


Figure 5.12: Results for total sample: Comparison of Projected Provisional Truth error rate ($E_{\mathrm{PT}(\theta)}$) v. projected reject rate ($F_{\mathrm{K}(\theta)}$) and minimum Projected Truth error rate: Method 2 ($\min E2_{\mathrm{T}(\theta)}$) v. maximum projected manual processing rate ($\max M_{\mathrm{T}(\theta)}$).

rate for the Production Data E_{PD}^{ref} as follows:

$$\begin{aligned} E_{\text{PD}}^{ref} &= (80\%)(1\%) + (20\%)(3\%) \\ E_{\text{PD}}^{ref} &= 1.4\% \end{aligned} \tag{5.1}$$

While the stated requirements applied strictly to the set of all write-in fields on all forms, we can apply them to the sample at hand for the purposes of evaluating PDQ performance. Let us say that as a rule of thumb, the reference Truth error rate E_{T}^{ref} should be a tenth of the reference Production Data error rate, and further, that the reference standard error σ_{T}^{ref} should be a tenth of the reference Truth error rate. We set our tuning parameters as follows:

$$D = 10 \tag{5.2}$$

$$E_{\text{T}}^{ref} = 0.14\% \tag{5.3}$$

$$d = 10 \tag{5.4}$$

$$\sigma_{\text{T}}^{ref} = 0.014\% \tag{5.5}$$

Given these parameter values, we can calculate the reference sample size f_{T}^{ref} using Equation 3.82 as follows:

$$\begin{aligned} f_{\text{T}}^{ref} &= \frac{(10)(10)^2}{1.4\%} - (10)^2 \\ f_{\text{T}}^{ref} &= 71,329 \end{aligned} \tag{5.6}$$

Additionally, we would like to have 95% confidence that our estimated Truth error rate is at or below the reference Truth error rate:

$$c = 95\% \tag{5.7}$$

$$z_c = 1.645 \tag{5.8}$$

Reviewing the weekly cumulative PDQ processing volumes, we find that the sample size for numeric fields on D-1(E) forms reached the reference sample size after Week 3 of May 2010. As shown in Table 5.2, the Truth error rate with the confidence interval applied ($E_{\text{T}} + 1.645\sigma_{\text{T}} = 0.00760\%$)

was lower than the reference Truth error rate ($E_{\mathrm{T}}^{ref} = 0.14\%$). Thus the conditions for tuning with the Projector model were satisfied.

The tuning chart in Figure 5.13 shows the relationships between the reference Truth error rate E_{T}^{ref}, Projected Truth error rate $\min E2_{\mathrm{T}(\theta)}$, and projected manual processing rate $M_{\mathrm{T}(\theta)}$, as functions of the confidence threshold θ as of Week 3 of May 2010. We wish to find the smallest threshold such that the Projected Truth error rate remains at or below the reference rate.

The critical values for tuning are shown in Table 5.3. A decision might have been made at that time to change the confidence threshold to 44. This would have increased the Truth error rate by a factor of 18 while keeping it below the reference rate. According to the model's predictions, the same decision would have decreased the manual processing rate to just over half the actual observed rate. At that point in time, PDQ had processed only about 2% of the total selected sample, so the future efficiency improvement would have been quite substantial.

For comparison, Figure 5.14 shows the tuning chart for the total sample as of the end of 2010 Census operations.

Table 5.4 shows some critical values. With the benefit of hindsight, we see that an earlier decision to set $\theta_0 = 44$ ultimately would have caused the Truth error rate ($\min E2_{\mathrm{T}(\theta)} + z_c\sigma_{\mathrm{T}(\theta)} = 0.20012\%$) to exceed the reference rate. In this case, we see that the "correct" tuning decision would have been to set the confidence threshold to 58, which would have decreased the manual processing rate to about 0.6 times the actual observed rate. A simple solution to this problem would be to continue monitoring a current subsample, approximately equal to the reference sample size, and to trigger a reset to the default, conservative confidence threshold ($\theta_0 = 80$) upon detecting that the Truth error rate was too high.

Table 5.2: Results through Week 3 of May 2010: Static model.

(a) Sample size.

Form Count	5,057
f_T	110,623

(b) Inputs.

E_{PD}	0.49447%
E_{PT}	1.05584%
E_{A1}	0.58845%
E_{A2}	0.58845%

(c) Probabilities, volumes, and Truth error rate.

y	$P[y]$	$F[y]$	$F[y] - P[y]$	$\lvert F[y] - P[y] \rvert$
PT=PD	98.45491%	98.45692%	0.00201%	0.00201%
A1=PD	1.04444%	1.04680%	0.00236%	0.00236%
A1=PT	0.48637%	0.48724%	0.00087%	0.00087%
A2=PD	0.00615%	0.00181%	-0.00434%	0.00434%
A2=PT	0.00286%	0.00000%	-0.00286%	0.00286%
A2=A1	0.00516%	0.00723%	0.00207%	0.00207%
INC	0.00011%	0.00000%	-0.00011%	0.00011%
				E_{T} = 0.00434%
				σ_{T} = 0.00198%
				$E_{\text{T}} + 1.645\sigma_{\text{T}}$ = 0.00760%

(d) Manual processing rate.

F_{K}	31.64532%
F_{A1}	1.54308%
F_{A2}	0.00904%
M_{T}	33.19744%

Table 5.3: Results through Week 3 of May 2010: Tuning values.

θ	E_{T}^{ref}	$\min E2_{\text{T}(\theta)} + z_c\sigma_{\text{T}(\theta)}$	$\max M_{\text{T}(\theta)}$
44	0.14%	0.13880%	17.50992%
80	0.14%	0.00760%	33.19744%

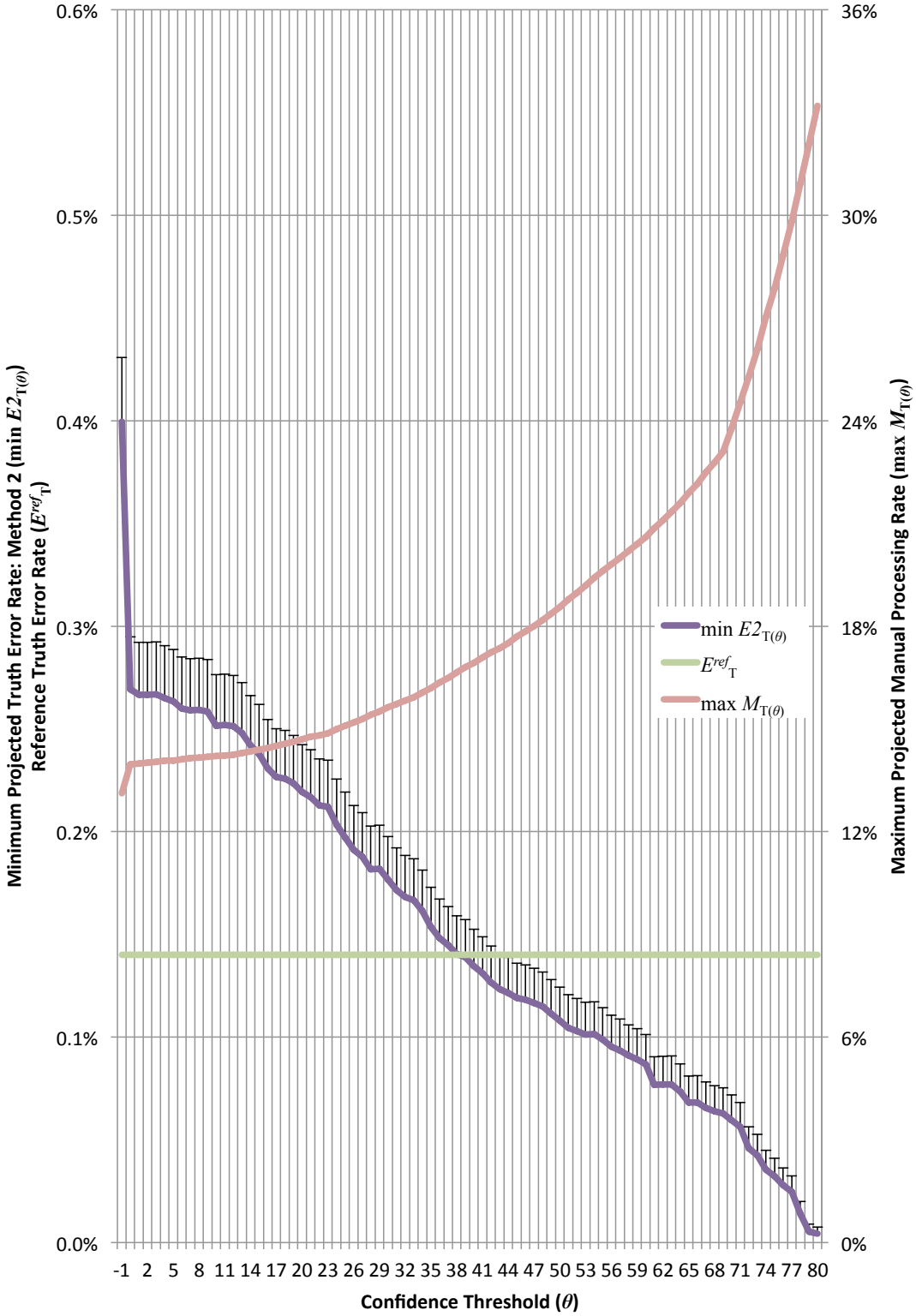


Figure 5.13: Results through Week 3 of May 2010: Tuning chart.

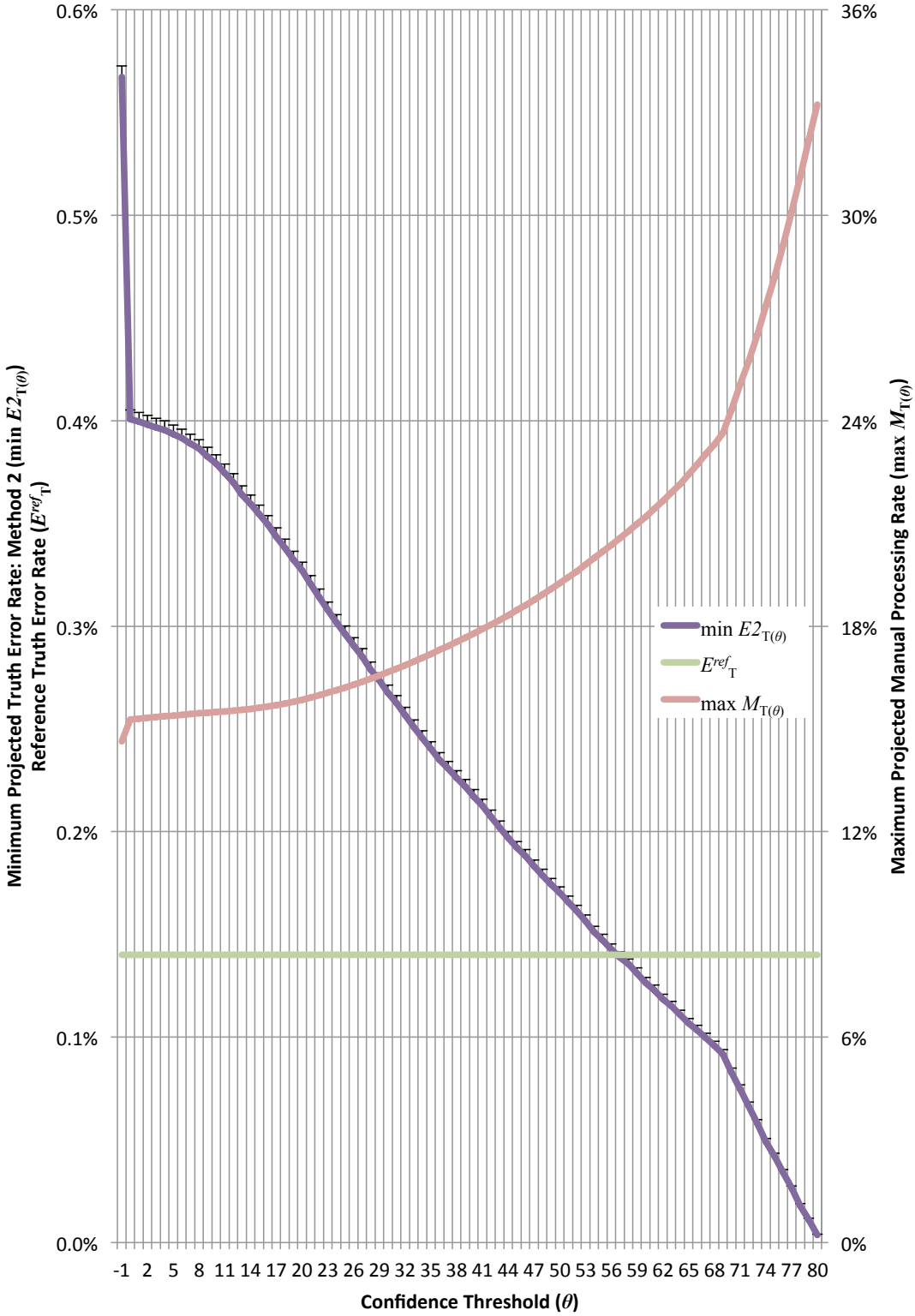


Figure 5.14: Results for total sample: Tuning chart.

Table 5.4: Results for total sample: Tuning values.

θ	E_{T}^{ref}	$\min E2_{\mathrm{T}(\theta)} + z_c\sigma_{\mathrm{T}(\theta)}$	$\max M_{\mathrm{T}(\theta)}$
44	0.14%	0.20012%	18.31843%
58	0.14%	0.13810%	20.77078%
80	0.14%	0.00399%	33.23876%

Chapter 6

Summary

6.1 Conclusions

PDQ has demonstrated its effectiveness in helping to ensure high quality production data capture for the 2010 Census. On a number of occasions, the static model has been used to validate the efficiency and Truth precision of the system. In some cases, we have attempted to tune confidence thresholds based on the Provisional Truth error rates and reject rates, with limited success. Most other efforts have focused on making major improvements to the capture quality of the Recognizer component. However, it is clearly desirable to have more complete understanding and control over the trade-off between the Truth error rate and the total manual processing rate. This motivated the Projector model, which I have developed and analyzed in this study.

For the selected sample, results from the static model showed an exceptionally low Truth error rate, almost 160 times as low as the Production Data error rate. However, this required manual processing of more than 33% of the fields. Using the results from the Projector model, we have seen how lowering the confidence threshold in the Independent Data Capture system can improve efficiency by sacrificing some degree of Truth precision. The respective minimum and maximum estimates of Projected Truth error rate by Methods 1 and 2 agreed quite well. This result suggests that the two different estimation methods are equally acceptable, and it also supports the validity of static model, upon which Method 2 is based. Seeing the complete dependence of both the maximum Projected Truth error rates on the

Projected Provisional Truth error rate, we conclude that the minimum estimates are better suited for practical use.

The Projector model, along with some heuristics for reference error rates and sample sizes, has allowed us to explore a realistic scenario for tuning PDQ's performance, without having to make substantial changes to the system's data capture processes. The results from a preliminary, minimal subsample showed the opportunity to reduce future manual processing by nearly half, while keeping the Truth error rate at an acceptable level. Further analysis of the complete sample showed that an initial decision would have been suboptimal in the long run, but overall there was still potential to reduce the manual processing cost by nearly 38%, or rather to increase manual processing efficiency by 60%.

6.2 Future Work

There are some known limitations to the Projector model; further elaboration of the model in these areas may yield additional insights. For example, in the calculations for $\min E2_{\mathrm{T}(\theta)}$ and $\max E2_{\mathrm{T}(\theta)}$, currently we do not identify which path-equivalent $y'(\theta)$ ultimately determines the estimated Projected Truth error rate. Also, currently we arrive at the estimates for Projected Truth error rate and projected manual processing rate independently of each other; that is, when we compute one of these perfomance metrics, we do not "track" our decisions about indeterminate outcomes in order to determine the consequent impact on the complementary metric. In addition, Method 2 for estimating the Projected Truth error rate assumes that E_{A1} and E_{A2} would remain fixed as the confidence threshold decreases; this is likely not the case in reality. Finally, if we view PDQ as a complex voting classifier, we may wish to examine how various internal factors in the Production Data Capture system, such as OCR confidence levels, impact PDQ's performance.

While we have made some strong inferences from the results of the Projector model, we do not have an operational instance of PDQ at hand to

confirm the predictions directly. A reasonable follow-up study might involve setting different confidence thresholds in the Independent Data Capture system and observing the outcomes. Among the outstanding questions are whether the minimum or maximum estimates given by the Projector model are closer to reality, and how those estimates and their components vary among different sample strata (by form type, field data type, etc.).

The tuning scenario described here involves realistic assumptions, and it should be possible to employ the methodology in future PDQ operations with just a few refinements. The selected sample for this study included all numeric fields from a particular form type. In practice, it may be useful to further stratify the sample by specific field groups, such as area code or birth year, as the subsamples grow sufficiently large; this would better account for any systemic data capture differences between these groups. In addition, if we have enough confidence in the Projector model's estimates, then there is potential to fully automate the tuning process. This can result in optimally efficient operation of PDQ with minimal expert intervention.

The current embodiment of PDQ is specifically geared toward paper forms data capture. However, the fundamental design, as described herein, is generically suited to many pattern classification domains, such as record linkage, fingerprint matching, or threat detection. The essential element is that in PDQ's Independent Classifier system (rather than Independent "Data Capture" system), the automated step (*e.g.*, Recognizer) associates both a decision (a response value, in the case of paper) and a confidence level with each unit of work (a field). This allows us to use the Projector model to understand the potential tradeoffs between Truth error and manual processing, and to make appropriate operational decisions.

Bibliography

[1] Advanced Document Imaging, LLC. Production Data Quality. `http://www.adillc.net/production-data-quality`, 2011.

[2] Michael Breithaupt. Improving OCR and ICR Accuracy Through Expert Voting: A White Paper. Technical report, Computing System Innovations, Apopka, FL, Dec 2006.

[3] C. K. Chow. An Optimum Character Recognition System Using Decision Functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, Dec 1957.

[4] C. K. Chow. On Optimum Recognition Error and Reject Tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, Jan 1970.

[5] Jay L. Devore. *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole—Thomson Learning, Belmont, CA, 6th edition, 2004.

[6] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., New York, 2nd edition, 2001.

[7] Douglass Huang. An Introduction to PDQ. Technical presentation, Version 9, Jul 2011.

[8] Emanuel Parzen. *Modern Probability Theory and Its Applications*. Wiley publication in mathematical statistics. John Wiley & Sons, Inc., New York, 1960.

[9] K. Bradley Paxton. Optimizing Paper Data Capture. *Today*, 29(5):26–28, Sep/Oct 2006.

[10] K. Bradley Paxton. Output Data Quality Criteria for PDQ. Technical report, Version 6, Jan 2009.

[11] K. Bradley Paxton. Paper Data Quality (PDQ) Keying Efficiency and Truth Precision. Technical report, Version 2, Dec 2010.

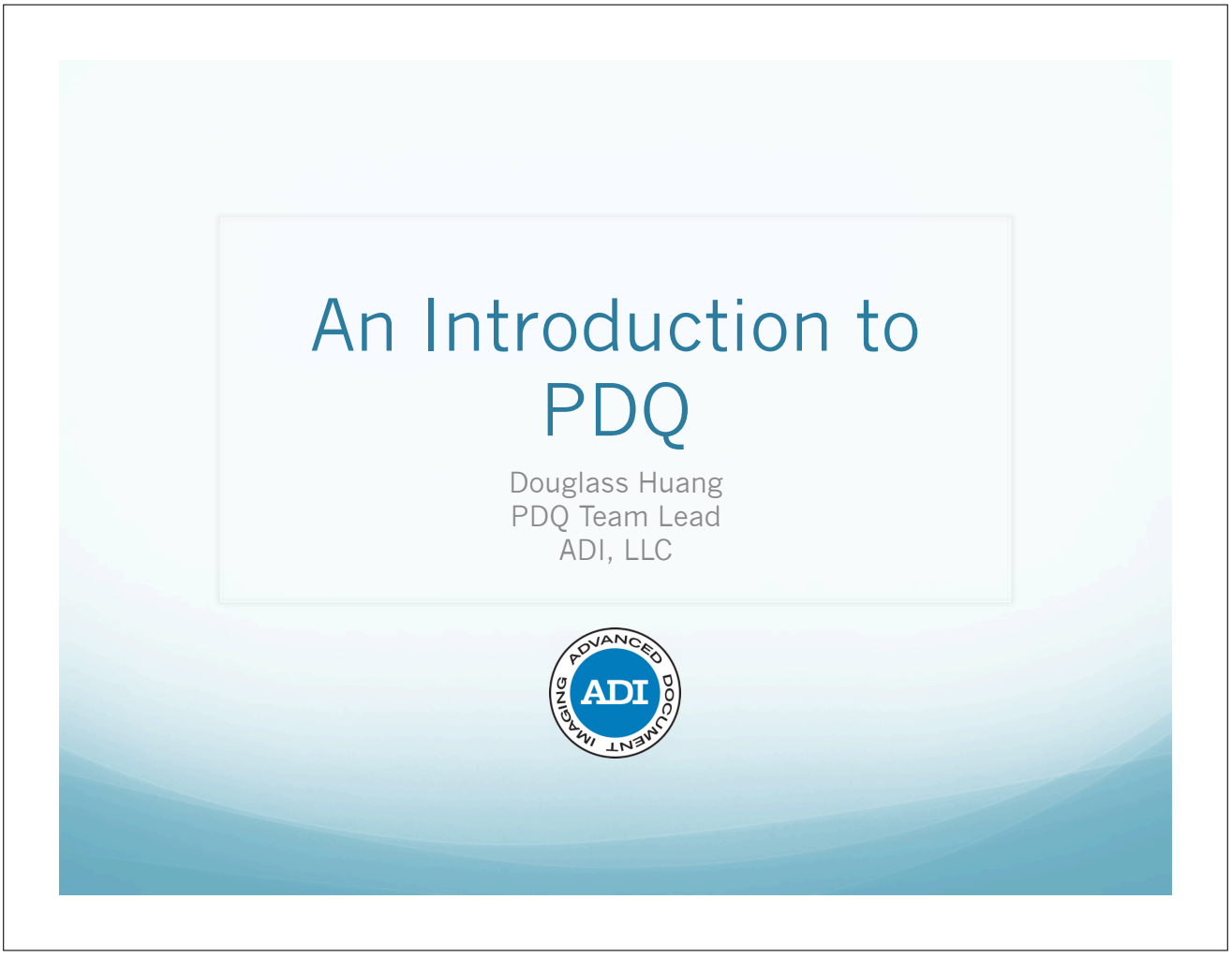
[12] K. Bradley Paxton, Steven P. Spiwak, and Douglass Huang. Truthing Production Data Capture. In *JSM Proceedings, Government Statistics Section*, pages 494–500, Washington, DC, 2009. American Statistical Association.

[13] U.S. Census Bureau. 2010 DRIS: RFP (Solicitation No.YA1323-05-RP-0006). `http://www.census.gov/procur/www/2010dris/dris-rfp.html`, Mar 2010.

Appendix A

Reproductions of Unpublished References

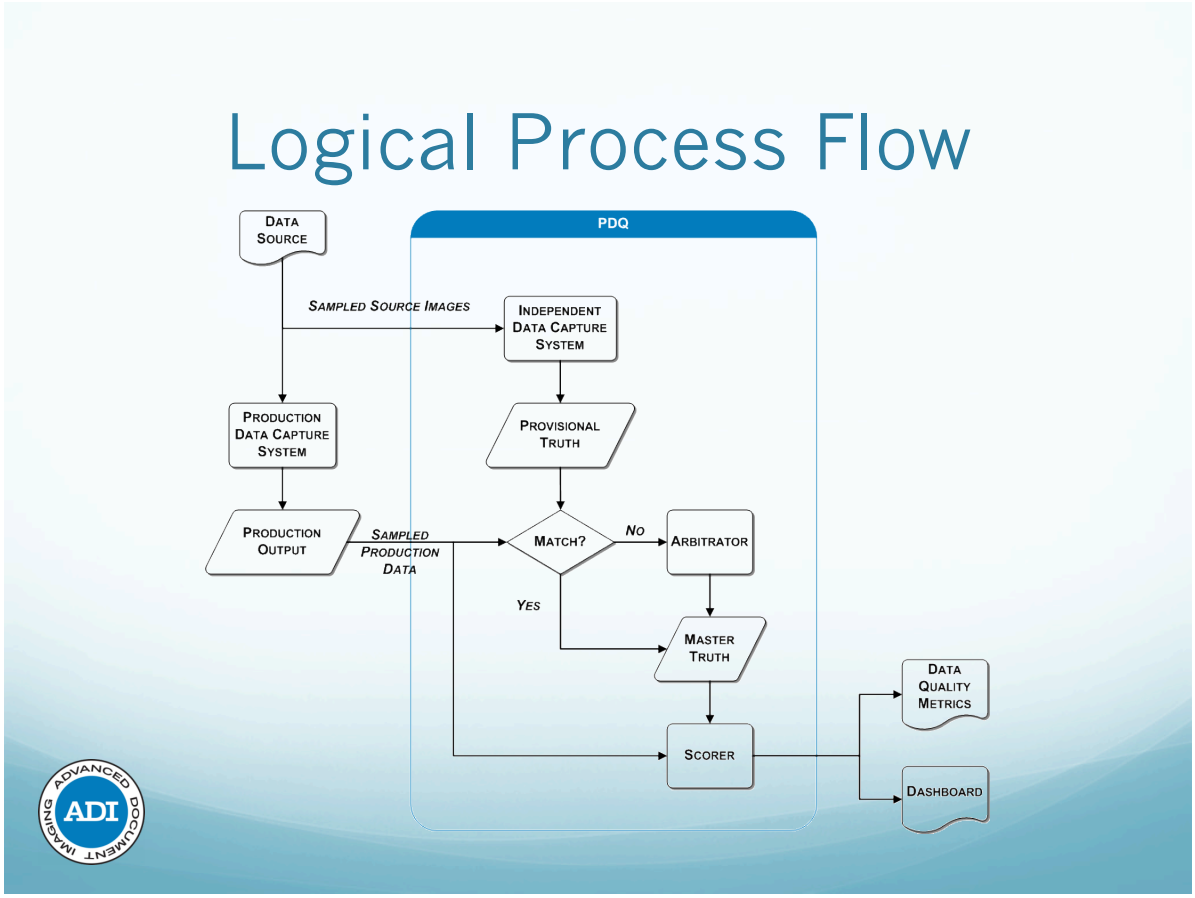
A.1 An Introduction to PDQ [7]



PDQ: Production Data Quality

Precise Efficient Timely Independent

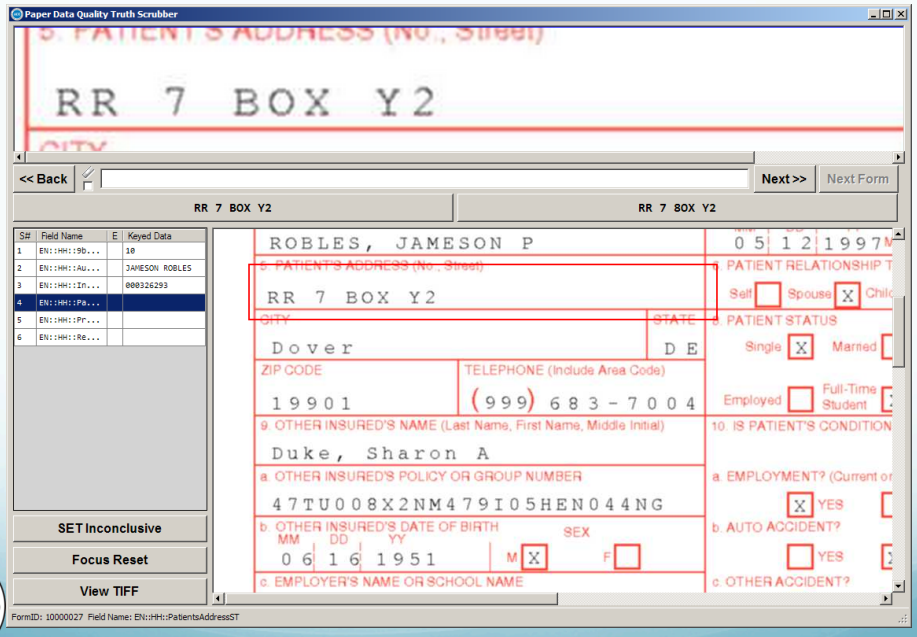
- Determines Truth of a sample of production source images with a high degree of precision
- Leverages software automation and sophisticated statistical design to minimize effort and operational cost
- When coupled with immediate sampling of production data, provides near-real time data quality measurements and feedback
- Employs independent OCR and OMR engines and expert human analysts to avoid systemic biases associated with internal verification processes


Logical Process Flow
PDQ
Data Source
Sampled Source Images
Independent Data Capture System
Production Data Capture System
Provisional Truth
Production Output
Sampled Production Data
Match?
No
Arbitrator
Yes
Master Truth
Data Quality Metrics
Scorer
Dashboard
ADVANCED IMAGING DOCUMENT
ADI

DRIS 2010 Outcomes

- PDQ processed a sample of 865K forms to assess the data quality of 164M Census forms
- PDQ process flow outcomes fit mathematical model closely, showing high efficiency and high precision
- Production Data vs. Provisional Truth Match Rate: 99.40%
- 97.46% of the 0.60% Arbitrator cases required only a single click by an Analyst to resolve the Truth
- Production Error Rate: 0.28%
- Estimated Master Truth Error Rate: 0.01%

"Truth Scrubber" Arbitrator Example



Dashboard

Intuitive Focused Flexible Extensible

- Intuitive user interface allows analyst to drill down via simple double-click
  - All the way down to the original image
- Focuses attention on pockets of errors occurring in the production data
  - Even for document types or field groups already meeting aggregate error rate requirements
- Supports advanced root cause analysis through built-in configuration, filtering, sorting, and *ad hoc* query capabilities
- Enables further analysis using popular off-the-shelf software packages via data export

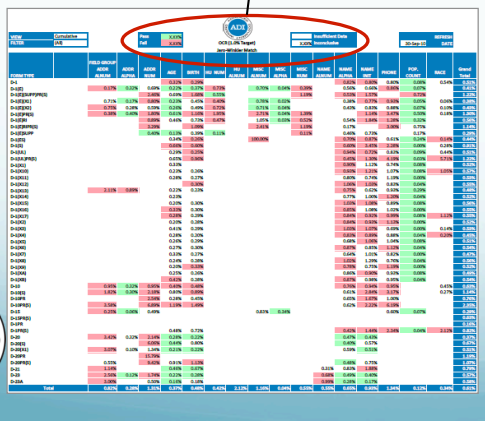
Dashboard Features

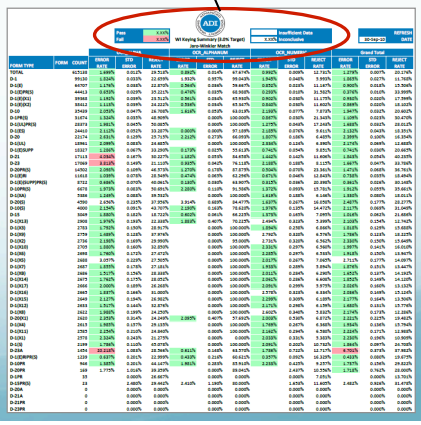
- Scoring with configurable matching function
- Pass/fail accuracy test
  - Aggregate-level display for monitoring overall compliance
  - Field-level display for root-cause analysis
- Weighting of aggregate-level results
  - Allows for different sampling rates per document type
- Processing history at the field level
  - Shows how production system arrived at its answer
  - Also includes PDQ field history
- Image Viewer

Dashboard Example

Pass	X.XX%
Fail	X.XX%

	Insufficient Data
X.XX%	Inconclusive





A.2 Output Data Quality Criteria for PDQ [10]

Output Data Quality Criteria for PDQ (Draft v.6)

Background
During a Paper Data Quality (PDQ) planning session between Census and ADI held in Rochester in December 2007, we discussed the types of data outputs to be coming from PDQ during production in Census 2010. There will be a lot of data contained in numerous planned reports, however Alan Berlinger asked the question of how would we be able to "flag" data quality problems in an effective and timely manner? We agreed to study that problem and make a recommendation to the Data Quality Integrated Project team (DQIPT), and, in particular, suggest possible "configurable parameters" that might be used in the software to achieve this goal. We reported on a way of doing this in March of 2008 with v. 5 of this paper. Based on discussions within the DQIPT, this new v. 6 is an update to simplify the outputs, and clarify when more samples are required to make firm conclusions.

Basic Data Quality Outputs from PDQ
The fundamental outputs from PDQ are error rates for both write-in fields and check-box fields on a number of form types. The DRIS data quality requirements have been documented, both by field type and form type. Essentially, however, they are to achieve 99% write-in field accuracy from the Optical Character Recognition (OCR) subsystem, 97% write-in field accuracy from the Key From Image (KFI) subsystem, and 99.8% field accuracy from the Optical Mark Recognition (OMR) subsystem. These are of course, obvious configurable parameters to start with, and we will just assume these are more or less correct for the purposes of this note.

Another important output from PDQ associated with a given class of data elements (e.g., Last Name fields on a DX-1 form), is the number of samples taken in the measurement. This data is needed in this analysis, as it affects our estimates of the sampling error, and, in turn, our confidence that we are meeting a particular requirement.

The Essential Approach
To begin with, we can use some simple approximate relationships to estimate the occurrence of data quality "problems". This is because we will only be looking to flag real problems that may occur in data quality. Also, because we expect the data quality to meet the requirements most of the time (based on our past experience), our approach generally is to look for strong evidence that a problem exists.

Example of a Measured Result

We use the letter q to represent an error, $\bar{q}$ to represent an average error measurement, and σ to represent the standard error associated with the sampling process, often called sampling error. An example of a measured result is shown below in Fig. 1.

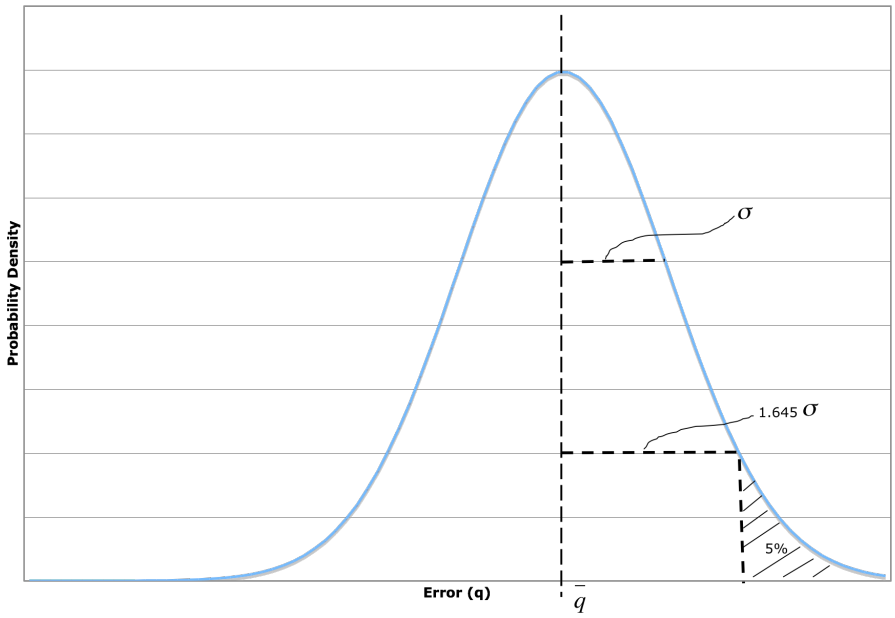


Fig. 1 – An Example of a Measured Error Distribution

Here, we have estimated an average error rate $\bar{q}$ by dividing the number of fields found to be in error by the number of samples in the measurement. The standard error, for our purposes, may be estimated from a simple binomial formula (discussed below), given the measured error rate, the sample size, and possibly, for small runner form types, the population size.

In this example, we show the "one-tail" estimate of confidence corresponding to 1.645σ above the mean, which says that there is an approximately 5% chance that the error rate in this measurement could be greater than $\bar{q} + 1.645\sigma$. Alternatively, we can say that we are "95% confident" that the actual system error is less than $\bar{q} + 1.645\sigma$.

Estimating the Standard Error
The binomial equation for sampling error σ is:

$$\sigma^2 = pq[1/n - 1/N]$$

where

σ = Standard (sampling) error
p = Success rate = $1 - q$
q = Error rate
n = Sample size
N = Population size.

The interesting thing about the above equation is kind of a "Catch 22", where you need to know the error rate to determine the standard error and decide if you have enough samples, etc. As a practical matter, after you actually measure the (mean) error rate, you can use that to estimate the standard error, for after all, it's the best estimate you have. (Note that if $N = n$, then $\sigma = 0$, since we sampled them all).

Since usually the population size N is very much greater than the sample size n, this equation simplifies (for large values of N relative to the sample size n) to:

$$\sigma^2 \approx pq/n$$

A practical approach to controlling sampling error is to define how large the sampling error should be relative to the error being measured, which we do with a parameter d, defined as:

$$\sigma = q/d$$

Our rule of thumb is that if d is ten or greater, that is a pretty good experiment as the standard error is less than $1/10^{th}$ of the error you are measuring. If d is around five, that's not bad for many practical applications. If it is around two, that's getting a little dicey, and if it is one, that is not a good result, and more samples are required.

We can now solve for the sample size in terms of p, q, and d, obtaining a result handy for estimating sample size:

$$n = pd^2/q$$

We can, for purposes of this paper, turn the last equation around, getting a way to estimate d from production data as:

$$d = \sqrt{nq/p}$$

We can use this result, and if $d < 2$, say, we may consider getting more samples of a particular form or field type during production.

Examples of Good, Fair, and Poor Outcomes

So we can estimate an average error rate, and using the sample size, estimate the associated standard error due to sampling. Below in Fig. 2 are shown three (of many) possible outcomes, which we have arbitrarily labeled Good, Fair and Poor relative to a required error rate, denoted by q_{req} and the vertical dotted line.

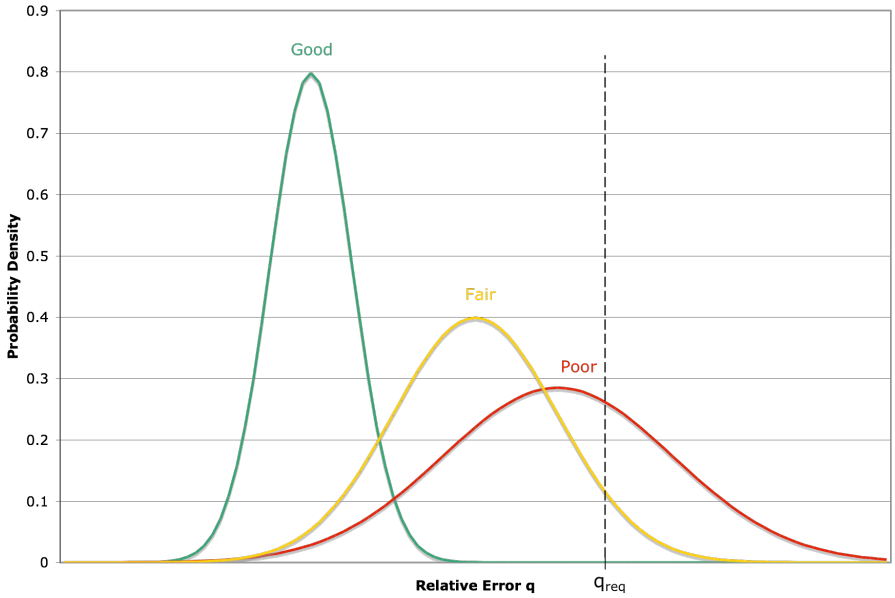


Fig. 2 – Examples of Good, Fair and Poor Measurement Outcomes

The green curve to the left is called "Good", because our mean measured error rate is significantly below the requirement, and the estimated standard error is small enough that we have a very low possibility that this error rate exceeds the requirement.

The yellow curve in the middle is labeled "Fair" because, even though the average error is less than the requirement, there is an approximately 5% possibility that the requirement is not met. (The data is probably all right, but we are becoming concerned, and would perhaps want to watch this one.)

The red curve on the right is labeled "Poor", because there is a significant probability, around 40% in this example, that our data quality is not meeting the requirement.

Therefore, in order to decide if we are content with data quality in a particular instance, we need to consider not only the average measurement, but also how confident we are that the measurement is correct.

Suggested Criteria
We could create quite a few criteria at this point, (and we suggested five relatively obvious criteria in the last version of this paper), however, the DQIPT felt that three were sufficient. Our recommended three criteria are expressed below in Table 1 by some qualitative descriptions, precise mathematical inequalities, and an approximate range of the probability of failure, by which we mean the probability that the data quality in a particular measurement does not meet the stated requirement.

Qualitative Description	Inequality	Approximate Range of Probability of Failure P_F
Good	$\bar{q} + 1.645\sigma < q_{req}$	$P_F < 5\%$
Fair	$\bar{q} + \sigma < q_{req} \le \bar{q} + 1.645\sigma$	$5\% \le P_F < 16\%$
Poor	$q_{req} \le \bar{q} + \sigma$	$16\% \le P_F$

Table 1 – The Three Suggested Data Quality Criteria

One could flag a data element as being "Poor" by simply waiting until we saw the average error exceed the requirement (as is often done), however, we have leaned in the direction of being more conservative here, to attempt to support Alan Berlinger's suggestion that we try and "spot problems earlier".

The internal process for "flagging" a data quality problem using PDQ is then:

1. Measure the mean error $\bar{q}$
2. Calculate the estimated standard error σ
3. Using the error requirement q_{req}, determine which of the above three inequalities is satisfied by the data in question (there can only be one)
4. Report the results in a convenient fashion.

The remaining concern is how to treat the case(s) where we don't have enough samples to make a firm conclusion. In order to handle this, we suggest we establish a threshold value of the parameter d (say, 2), and if d is less than this value, we qualify our results accordingly, and attempt to obtain more samples for that case.

The results might be reported in a display screen as shown in the next section.

A Data Quality Display Idea
It seems that this approach might lend itself to a visual display that could be easy for Census Management to graphically "see how things are going". Of course, another configurable parameter might be the latency of the results display. As an example, we might display the previous day's results on one screen, and a running average on another. For now, we assume that daily postings are sufficient.

The screen might look something like illustrated in Fig. 3 below, which could be created, for example, using the configurable formatting feature of Excel or other suitable means:

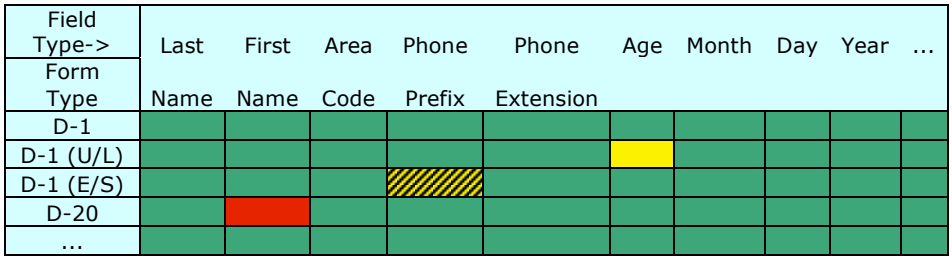


Fig. 3 – A Data Quality Display Idea

The above could be just the upper left-hand corner of a larger display, say a large plasma or LCD flat screen. We are summing up the data for like field types here (i.e., Last Names, as opposed to Last Name 1, etc.), to keep the number of columns in the display feasibly small.

In this example, we see most of the data quality is coming out good (as expected), represented by Green. There is a concern beginning to develop for the Age field on the D-1 (U/L) form, represented by Yellow (fair), and we should keep our eyes on that one going forward. We see a potential problem emerging for the Phone Prefix fields on the D-1 (E/S) form, but it is crosshatched, indicating that PDQ should try and obtain more samples of that data element. There is a problem of poor quality shown for the First Name field on the D-20 form represented by Red that deserves investigation by the appropriate data capture process engineers.

Configurable Parameters
A short list of the configurable parameters that would have to be specified (by form type and by field type) to produce this sort of data quality display for Census Management is given below:

- OCR Write-In Field Accuracy Requirement (High Confidence Fields)
- KFI Write-In Field Accuracy Requirement
- OMR Check-Box Field Accuracy Requirement
- Number of Fields in Sample
- Population (for short runner forms?)
- Display Latency (assume daily for now)
- Threshold value of parameter *d* to indicate more samples are needed
- The desired "Z" statistic (i.e., the value of 1.645 used above)

Conclusions
A simple approach to "flagging" potential Census 2010 data capture quality problems from PDQ in a timely and cost-effective manner has been described. A short list of configurable parameters is given to enable the process. This scheme should be a useful tool to help Census management monitor data quality during the 2010 Census data capture production.

ADI, LLC/kbp
12 Jan 2009

A.3 Paper Data Quality (PDQ) Keying Efficiency and Truth Precision [11]

Paper Data Quality (PDQ) Keying Efficiency and Truth Precision

Background

Emerging from it's early beginnings in measuring data capture quality during the Year 2000 Decennial Census, the PDQ system has evolved into a very efficient and precise way to determine the truth of a data capture operation's output. Once this truth is determined, the production data quality may be measured, problems identified, and system improvements made and verified in a cost-effective manner.

Prior to the creation of PDQ, (and still very common today), the standard procedure for determining the truth of production data capture operations was to manually perform what is called Double Key and Verify (DK&V), which is costly, time-consuming, and prone to error.

The PDQ system was used to independently measure the data quality of the Year 2010 Decennial Census, and verify that the quality metrics for the Decennial Response Integration System (DRIS) met the agreed requirements set by the Census Bureau. These metrics were the accuracy of Optical Character Recognition (OCR), Optical Mark Recognition (OMR), and Key From Image (KFI). Approximately 800,000 forms were processed by PDQ during the 2010 Census, and this collection of data may be the largest database of production Census output for which the truth is known existing today. In addition, many small pockets of error were uncovered during production in a timely fashion, allowing for reprocessing of certain forms to achieve better final results.

Summary

In the analysis that follows, we show that the keying efficiency of PDQ relative to DK&V is over 28 times, that is, for a given test sample of production forms, the quality assurance keying effort using PDQ is 28 times less than for DK&V.

This means that the same amount of test sampled forms can be processed with $1/28^{th}$ of the QA keying staff, or that twice as many samples can be processed with $1/14^{th}$ of the QA keying staff, etc., to suit the client's need.

It is also shown that the resultant truth error using PDQ is from 6 to 10 times less than DK&V, depending on the quality of the keyers.

Further, the efficiency of PDQ can be increased from 28X up to 40X depending on the particular application test requirements.

Analysis – Keying Efficiency

In order to derive metrics for PDQ efficiency and truth precision, we use a collection of F = 78,122 D-1 "short" forms processed by PDQ during the 2010 Decennial Census. This is a reasonable choice for our test universe, because the D-1 form constituted over half of the total number of forms processed by DRIS, and the data therein is typical Census data.

These forms contained a total number of f = 5,385,747 fields, both write-in and check-box. This about f/F = 69 fields per form.

All of these forms (and fields) were processed by PDQ to determine the truth, and allow data quality scoring. If someone were to use DQ&V to determine the truth they would have to key all the fields twice, or $2f$ = 10,771,494 fields.

The PDQ system employs automation as well as sophisticated statistical design to do much less keying than for standard DQ&V. The first portion of keying is reject keying for the independent recognition subsystems, which we call Form Completion, that is f_{FC} = 348,045 fields. The second portion is a smaller effort we call Truth Scrubber, wherein f_{TS} = 32,854 fields were keyed. The total fields keyed by PDQ then for this test universe of 78,122 D-1 forms was $f_{PDQ} = f_{FC} + f_{TS}$ = 380,899 fields.

Therefore, the efficiency factor for PDQ keying relative to DK&V is $2f/f_{PDQ}$ = 28.3.

We are presently examining the internal trade-off within PDQ between Form Completion and Truth Scrubber to see if greater efficiency can be obtained. A current rough estimate says the efficiency can be increased to 40X or more, depending on conditions.

Analysis – Truth Precision

In Reference 1 is a detailed derivation of a probability model for PDQ, and some data showing how well the model and the data coincide. That data was based on the 2008 Census Dress Rehearsal, and was a preliminary step in the development of PDQ.

Recent data and model comparisons based on actual 2010 Census production data were recently made available, and shows that the system has been improved and that the model fit the data extremely closely. Below is a table describing the model and data for our D-1 universe for both write-in and check-box fields:

W-I + Ck-box			
Probability	Theory	Data	Data-Theory
P_h	99.39650%	99.40454%	0.00804%
P[K_1=DRIS]	0.31605%	0.30956%	-0.00649%
P[K_1=PDQ]	0.27633%	0.27079%	-0.00554%
P[K_2=DRIS]	0.00536%	0.00336%	-0.00200%
P[K_2=PDQ]	0.00469%	0.00169%	-0.00300%
P[K_2=K_1]	0.00088%	0.00951%	0.00863%
P[I]	0.00020%	0.00056%	0.00035%

Without getting too wrapped-up in the details (see Reference 1), the first row of numbers in the above table refers to the probability that a hard match occurs between the production data file and the PDQ Provisional Truth file, which you will note is very high, about 99.4%. This is almost all of the data, and means that only 0.6% of the fields being processed by PDQ move on to the Truth Scrubber step.

The difference between the theory and the data is a very small 0.008%. Since this row accounts for most of the data, and since the probability theory does not account for truth error (but, of course, the data may contain some), then it follows that the 0.008% is an (upper bound) estimate for truth error in PDQ.

For DK&V, if the keying error is e_k then there would be about $2e_k$ errors to be verified by a person who also may have an error roughly equal to e_k, and so an estimate for DK&V truth error is about $e_{DK\&V} = 2e_k^2$. If keyer error e_k is about 2%, then $e_{DK\&V} = 0.08\%$, and that assumes everything else is done correctly. (By correctly, we mean that there is no collusion between keyers, that they follow the keying rules consistently, etc.) If the keyers were better, say, more like DRIS keyers were during production and $e_k = 1.5\%$, then $e_{DK\&V} = 0.05\%$.

Therefore, the PDQ truth error is about 6 to 10 times smaller than DK&V, depending on DK&V keyer quality.

References

1 Paxton, K. Bradley, Spiwak, Steven P., and Huang, Doug, 2009, *Truthing Production Data Capture*, Joint Statistical Meetings (JSM) Proceedings, Government Statistics Section, Washington, D.C.: American Statistical Association, 494-500.

ADI, LLC/kbp
20Dec2010