Downloaded from UvA-DARE, the institutional repository of the University of Amsterdam (UvA)
http://hdl.handle.net/11245/2.176229

File ID uvapub:176229
Filename Thesis
Version final

SOURCE (OR PART OF THE FOLLOWING SOURCE):
Type PhD thesis
Title Measurement system analysis for binary tests
Author(s) T.S. Akkerhuis
Faculty FEB: Amsterdam Business School Research Institute (ABS-RI)
Year 2016

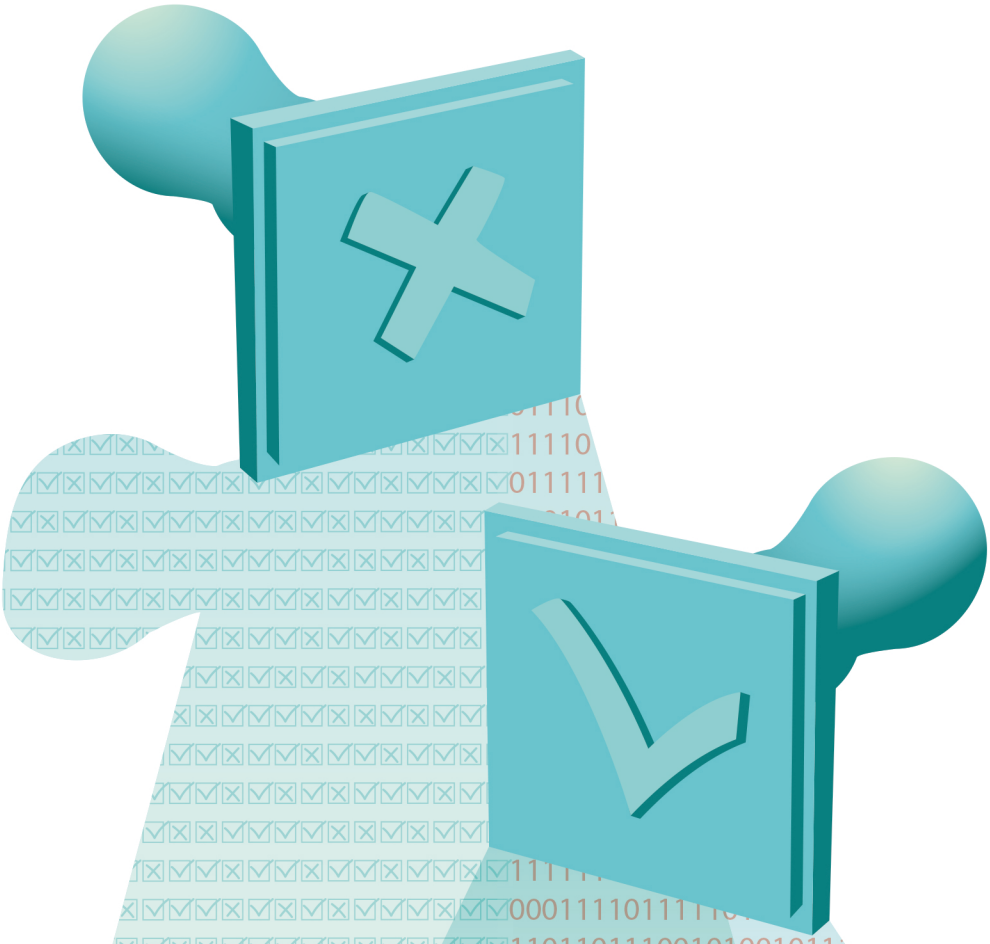
FULL BIBLIOGRAPHIC DETAILS:
http://hdl.handle.net/11245/1.540065

Copyright

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content licence (like Creative Commons).

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
(pagedate: 2016-08-31)

Measurement System Analysis for Binary Tests



Thomas Akkerhuis

Measurement System Analysis for Binary Tests

Thomas Akkerhuis



Instituut voor Bedrijfs- en Industriële Statistiek

Publisher IBIS UvA, Amsterdam - www.ibisuva.nl
Printed by Gildeprint, Enschede - www.gildeprint.nl
Cover design Remco Wetzels, Amsterdam - www.remcowetzels.nl
ISBN 9789462333673

Measurement System Analysis for Binary Tests

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 16 september 2016, te 14:00 uur

door

Thomas Simon Akkerhuis

geboren te Grubbenvorst

Promotiecommissie

Promotores

Prof. dr. J. de Mast	Universiteit van Amsterdam
Prof. dr. M.R.H. Mandjes	Universiteit van Amsterdam

Overige leden

Prof. dr. P.M.M. Bossuyt	Universiteit van Amsterdam
Prof. dr. R.J.M.M. Does	Universiteit van Amsterdam
Prof. dr. M. Salomon	Universiteit van Amsterdam
Prof. dr. S.H. Steiner	University of Waterloo
Dr. E.G. Bashkansky	ORT Braude, Israël
Dr. M. Schoonhoven	Universiteit van Amsterdam
Dr. W.N. van Wieringen	Vrije Universiteit

Faculteit
Economie en Bedrijfskunde

Contents

1 Introduction **1**
- 1.1 A conceptual model of measurement . . . 3
- 1.2 Notation and context . . . 7
- 1.3 Literature review . . . 8
- 1.4 Objective and outline of this thesis . . . 10
- 1.5 Contributions to scientific literature . . . 11

2 A mechanistic model for binary measurements **13**
- 2.1 Introduction . . . 13
- 2.2 Methodology . . . 15
  - 2.2.1 Statistical model . . . 15
  - 2.2.2 Sampling strategy and estimation . . . 18
  - 2.2.3 Model diagnostics . . . 21
- 2.3 Suitable sample sizes and robustness . . . 22
  - 2.3.1 Optimal proportions of sample sources . . . 22
  - 2.3.2 Precision as a function of I^{Rej} and K . . . 23
  - 2.3.3 Robustness to misspecification . . . 24
- 2.4 MSA Study at a car parts manufacturer . . . 26
  - 2.4.1 Background of the MSA study . . . 26
  - 2.4.2 Statistical analysis . . . 27
  - 2.4.3 Discussion and appraisal of the car parts MSA study . . . 31
- 2.5 Discussion . . . 32

3 Estimation of *FAP* and *FRP* without gold standard: mathematical analysis **35**
- 3.1 Introduction . . . 35
- 3.2 Statistical modeling . . . 37
  - 3.2.1 Rejection counts U_i and rejection probabilities R_i . . . 37
  - 3.2.2 Motivation by a mechanistic model . . . 38

3.2.3 Estimation of FAP and FRP is problematic . . . 39
3.3 Estimation of random error . . . 43
3.3.1 Estimation of IAP and IRP . . . 45
3.3.2 Estimation of IAP and IRP in the presence of nuisance variables 46
3.3.3 Design of the MSA study . . . 48
3.4 Conclusions . . . 49

4 Estimation of FAP and FRP without gold standard: simulation study 51
4.1 Introduction . . . 51
4.2 Theory and methods . . . 52
4.2.1 Statistical modeling . . . 53
4.2.2 Methods included in the comparison . . . 55
4.2.3 Scenarios . . . 58
4.2.4 Implementation details . . . 58
4.2.5 Simulation details . . . 59
4.3 Findings . . . 61
4.3.1 Traditional latent class method . . . 61
4.3.2 Approaches based on characteristic curves . . . 64
4.3.3 Latent class random effects method . . . 64
4.3.4 Estimation in the presence of nuisance variables . . . 66
4.4 Discussion and conclusions . . . 66

5 Robust estimation of IAP and IRP 69
5.1 Introduction . . . 69
5.2 Statistical modeling . . . 70
5.3 Estimation of f_R^{Tot} by adaptive polynomials . . . 73
5.4 Quantitative evaluation by means of simulation . . . 75
5.4.1 Robustness . . . 76
5.4.2 Sample size recommendations . . . 79
5.5 Example: optical inspection of car parts . . . 81
5.6 Discussion and conclusion . . . 83

6 Probabilities of inconsistent classification for nominal measurements 85
6.1 Introduction . . . 85
6.2 Statistical modeling . . . 87
6.2.1 Modeling of measurement outcomes . . . 87
6.2.2 Estimation of ICPs . . . 89
6.2.3 Diagnostics and tests . . . 90
6.2.4 Probabilities of mis- and inconsistent classification . . . 92
6.3 Evaluation by means of simulation . . . 93

6.3.1 Effect of different concentration β . . . 93
6.3.2 Effects of non-uniformities in $\boldsymbol{\alpha}$ . . . 94
6.3.3 Effect of number of classes C . . . 95
6.3.4 Effect of sample size I and number of repeated measurements K 95
6.4 Real-life case . . . 97
6.4.1 Background . . . 97
6.4.2 Results . . . 98
6.4.3 Probabilities of inconsistent classification . . . 100
6.4.4 Hypothesis testing . . . 100
6.4.5 Diagnostics . . . 102
6.4.6 Conclusions from the case study . . . 103
6.5 Conclusions . . . 103

7 Summary **105**

Bibliography **109**

Samenvatting (Summary in Dutch) **115**

Curriculum vitae **119**

Acknowledgements **121**

Chapter 1

Introduction

Commercial banks have operations called cash centers, where retailers, bars and restaurants deposit cash money (in particular bank notes). The cash centers count this money, credit the owner's bank account, and prepare the counted bank notes for recirculation. In addition, the counting machines check bank notes for authenticity. Based on whether required authenticity features are detected, bank notes are either classified as "rejected" (possibly forged) or "accepted". This classification aims to reflect whether notes are truly authentic or forged. This thesis is about the statistics behind such classifications on two-point ("binary") scales, and we consider them as a form of measurement. They have in common with numerical measurements that the classifications aim to reflect a true property of the measured items ("forged" or "authentic"). Binary measurement differs from numerical measurement in that the outcomes are not numbers, but one of two classes such as "accept" or "reject".

Unfortunately, the inspection machines make errors. Notes can be falsely rejected due to folding, dirt or regular printing variations, or they can be falsely accepted. Both types of mistakes are costly. Rejects are typically reinserted, or inspected manually, so that the process cannot be automated. False acceptance is a direct loss to the cash center. An event in which an outcome does not reflect the true state is called *measurement error*. For evaluating the performance of competing inspection machines, the rates or probabilities of a false rejection or a false acceptance could be determined. Such an evaluation is an example of *measurement system analysis* (MSA). MSA is the branch of statistics that is concerned with designing experiments, models and analyses for evaluating measurement systems.

How to perform an MSA study then? One can take a sample of bank notes, and have them checked for authenticity by both the inspection machine and an expert. Assuming that the expert's judgment is infallible, we can determine the probability of a false rejection and the probability of a false acceptance by comparing the judgments

of the machine to those of the expert. Thus, the expert's judgments are used as a *gold standard* (also called *reference standard*). Unfortunately, that will not work in this case, as some authenticity features are kept secret by the national bank, and no expert is infallible. That means that the bank notes' *true values* (also called *true states*) are unknown. The determination of the inspections' false rejection and false acceptance probabilities then becomes a much more difficult pursuit. It is this situation that we consider in this thesis, and we refer to it as the *gold standard unavailable situation*. The thesis's objective is to develop MSA approaches for binary measurements in such situations where a gold standard is not available.

The importance of binary MSA is not limited to the inspection of bank notes. For example, in medicine, diagnostic tests give "positive" or "negative" outcomes that aim to reflect the true presence or absence of a medical condition, and *measurement error* can result in waste due to unnecessary treatment, or worse, endanger a patient's health. In industry, quality inspections judge whether products meet their specifications and yield "accept" or "reject", and errors increase costs of complaints handling and loss of goodwill, or unnecessary scrap or rework. By means of a conviction or acquittal, a judge aims to reflect whether the defendant has committed a crime or not, but he or she can be mistaken. Type I and type II errors are well-known concepts in statistics that refer to unjust rejection or non-rejection of statistical hypotheses.

Also the unavailability of a gold standard is not limited to the example of authenticity of bank notes. A gold standard is often undefined (or underdefined) when it comes to perceptual qualities of products (for example, whether a product's taste is good or bad), or when a reference or standard is simply unknown (an example will be presented in Chapter 2). Application of the gold standard test can also be invasive or even destructive and for that reason considered unavailable: an MRI-scan is often used to detect brain tumors, but the gold standard is a brain biopsy. Finally, a gold standard may be unavailable due to financial reasons, for instance when it involves costly laboratory analysis.

Binary measurement systems are poorly understood, as are the consequences of the unavailability of a gold standard. One cause is the treacherous simplicity of the measurement scale: an inspection machine needs to measure countless properties from a bank note, all of which are collapsed into a simple accept/reject decision. Current methods are motivated by models that make overly simplistic or otherwise unrealistic assumptions, and therefore may lead to quantifications of measurement error that are substantially biased. The objective of this thesis is to understand these shortcomings, and to find a method that does not depend on such restrictive assumptions. In the next sections, we will formally define measurement and measurement error. Further, we give an overview of research on binary MSA, and how the chapters of this thesis contribute to the current literature.

1.1 A conceptual model of measurement

When we drive on the highway, from time to time we are overtaken by another car. This is an empirical relation between both cars with respect to the property speed: the other car has a higher speed than we have. This relation can also be expressed mathematically, by assigning symbols (in this case, numbers) to the two cars, reflecting speed. We might find 95 kph for us, and 100 kph for the other car. The empirical relation that the other car drives faster is then reflected by the mathematical relation $100 > 95$. A formal definition of measurement is: the assignment of symbols to items (things, subjects, or other phenomena) in such a way that mathematical relations among the symbols represent empirical relations between the items with respect to a property under study. This definition comes from a branch of mathematics and philosophy of science called *measurement theory* (Wallsten, 1988; Hand, 1996). Similar definitions are generally accepted in the social sciences (Allen and Yen, 1979; Lord, 1980). We elaborate on this definition of measurement as a mapping from a set of items to a set of symbols, to explain in what way binary classifications can be understood as a form of measurement. This understanding provides us with a solid conceptual basis allowing us to define the notions of true value, gold standard and measurement error for binary measurement. This in turn allows us to give a precise definition of the topic and objective of the thesis.

Measurement is a mapping from a set of items to a measurement scale. The latter is a set of numbers or symbols, such as $\mathbb{R}^+$, $\{0, 1\}$ or $\{A, B, C, D\}$, equipped with algebraic structures and operators such as $\leq$ (order) and $+$ (addition). Measurement constitutes a homomorphic map, meaning that it preserves structure. We explain what this means by considering a set of physical objects $\mathcal{I}$ and one of their properties, namely, their mass. This empirical property induces various relational structures among the objects, such as order. One can, for instance, compare two objects $A, B \in \mathcal{I}$ directly using a balance (i.e., without measuring them) and establish that one of them (say, B) is heavier ($A \sqsubseteq B$). Even richer structures are created if we allow operations applied to the items. For example, let us denote by $A \circ B$ the operation of grouping two objects together on one side of the balance, and let $A \circ B \sqsubseteq C$ denote the empirical finding that a balance with A and B on one side and C on the other tips down on the latter side. By $[\mathcal{I}, \sqsubseteq, \circ]$ we denote the empirical system that consists of a set objects $\mathcal{I}$, and the relational structures induced by the empirical ordering $\sqsubseteq$ and the grouping operation $\circ$.

By measuring the mass of the objects, using a spring scale, for example, we map these empirical relations to a set of numbers equipped with mathematical structures. The measurement assigns to each measured object $A \in \mathcal{I}$ a numerical value $M(A) \in \mathbb{R}^+$ and, thus, measuring in this case is a map $M : \mathcal{I} \to \mathbb{R}^+$. The set of numbers $\mathbb{R}^+$ is equipped with mathematical relations, such as the order relation $\leq$, and the idea of

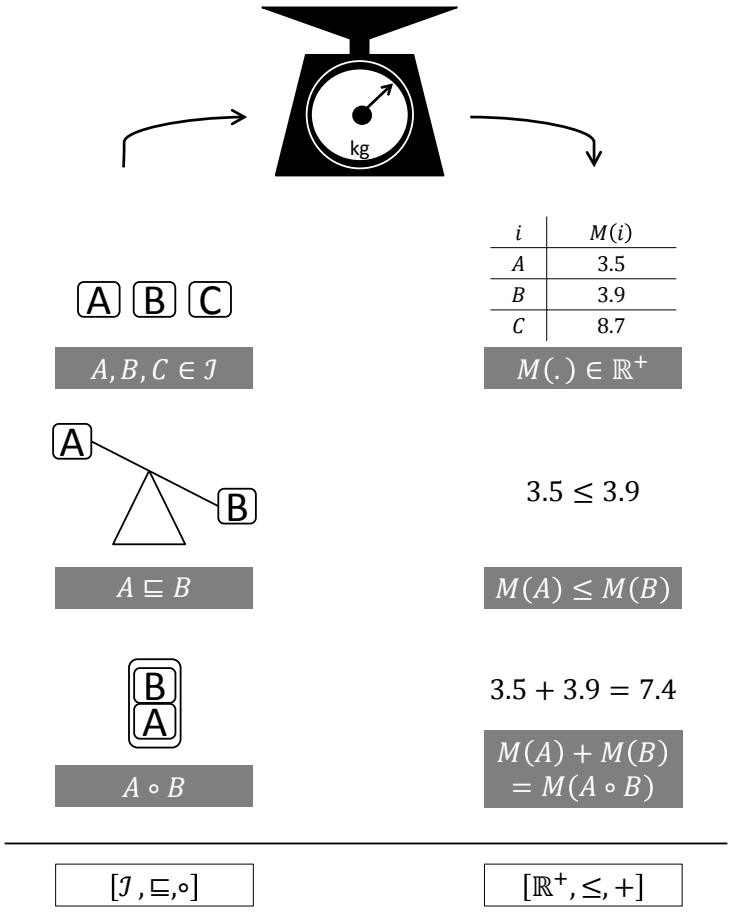


Figure 1.1: Measurement of mass is a homomorphic map $M : \mathcal{I} \to \mathbb{R}^+$, which preserves order (differences $\sqsubseteq$ established by a balance are mathematically represented by inequality $\leq$) and addition (grouping together of mass $\circ$ is mathematically represented by addition $+$).

measurement is that the mathematical order between measured mass values reflects the empirical order given by the comparison of objects using a balance (that is, if $A \sqsubseteq B$ then $M(A) \leq M(B)$). Moreover, $\mathbb{R}^+$ is equipped with algebraic operators such as $+$ (addition), and the idea of measurement is that they have empirical counterparts such as the grouping of objects together on one side of the balance ($A \circ B$) and that relations carry over ($M(A \circ B) = M(A) + M(B)$). Measurement, therefore, maps a set of items and its structures such as $[\mathcal{I}, \sqsubseteq, \circ]$ onto a set of measurements with structures such as $[\mathbb{R}^+, \leq, +]$. The mathematical term for such structure preserving map is a *homomorphism*. This homomorphism ensures that mathematical statements such as $M(A) + M(B) \leq M(C)$ have empirical meaning, in the sense that we observe that C has greater mass than A and B combined. See Figure 1.1 for a graphical representation of a homomorphic map.

Some measurements preserve more structure than others, meaning that the measurement scale allows for a richer set of mathematical operations. An influential typology of measurement as to how much structure it preserves is the one by Stevens (1946). He discerns the following levels of measurement:

- *Nominal* measurement preserves equivalence relations ($=$) only. This amounts to mere categorization. An example from industry is the assignment of failure modes to products that failed quality inspection.
- *Ordinal* measurement preserves equivalence and order relations ($=$ and $\leq$). The categories of an ordinal scale have a defined order, such as quality inspections in terms of "reject", "acceptable" or "excellent".
- *Interval* measurement preserves, in addition, distances between items and thus allows addition and subtraction.
- *Ratio* measurement preserves, in addition, an empirical zero point and thus allows multiplication and division. The difference between interval and ratio measurement will be explained in more detail below.

Other types of measurement can be defined in terms of this framework as well. Binary measurement is the special case of nominal or ordinal measurement on a scale with only two classes. Categorical measurement is measurement on either of the first two levels, and numerical measurement is measurement on either of the second two levels. The distinction between *continuous* and *discrete* is useful for random variables but not for measurement data, because these are always discrete. Next we discuss some of the practical ramifications of the level of a measurement scale.

$\mathbb{R}^+$ is equipped with order ($\leq$), a norm or distance metric (e.g., 3 and 1 are twice as far apart as 2 and 1), and a zero point. When measuring mass in kilograms or pounds, each of these has an empirical counterpart (for example, $M(A) = 0$ corresponds to

the total absence of mass). But when one measures temperature in degrees Celsius or Fahrenheit, one uses the same numerals in $\mathbb{R}$, but some mathematical relations and algebraic operators do not have empirical counterparts. In particular, the zero point of neither of these temperature scales has empirical meaning in the sense that it corresponds to the total absence of an empirical quantity. As a consequence, a statement such as: "30 °C is two times 15 °C" is true about the numbers themselves, but it is difficult to see what empirical meaning such statement has. The Celcius and Fahrenheit scales are interval scales, and scalar multiplication and division have no empirical meaning. Note that the Kelvin scale does have a zero point that corresponds to the total absence of the empirical quantity in question and, consequently, "30 K is twice as warm as 15 K" does have empirical meaning. Temperatures expressed in Kelvin are thus on a ratio scale.

On an ordinal scale such as $\{A, B, C, D\}$, only the $\leq$ and $=$ relations have empirical meaning. However, there is no distance metric, and operators such as addition and subtraction do not in general have empirical meaning. Recoding the scale using numerals (that is, $1 = A, 2 = B, 3 = C, 4 = D$), one could apply mathematical operators to these numerals. Still, a statement such as: "the difference between C and A is twice as large as the difference between B and A" (since $3 - 1$ is two times $2 - 1$) is in general empirically meaningless. By implication, also statistics based on addition and subtraction (such as the average or standard deviation) are typically meaningless for ordinal data.

The conception of measurement as a homomorphism shows that, in addition to the measurement of quantitative properties such as temperature and weight, also ratings on a nominal scale, quality inspections on a binary scale, and diagnoses by radiologists or physicians on a nominal or ordinal scale can be conceived as measurements. Thus, the given definition seems more general than typical definitions of measurement in metrology (Kimothi, 2002; JCGM, 2008) that seem especially geared to quantitative measurements (JCGM, 2008, p.16, explicitly states that the word measurement does not apply to nominal properties).

In addition to the notion of measurement as a homomorphism, we need the concepts of measurand, true value, reference value, and measurement error. The *measurand* is the empirical property that the measurements aim to reflect. It can be a continuous property, such as the mass of objects, but also a dichotomous or polytomous property or a more complex combination of properties. The definition of the measurand should not be confused with the operational definition of the measurement procedure. Often, a measurand is not assessed directly but via a known relationship with other, more readily observable properties. Examples include the determination of an object's mass by determining the compression of a spring in a scale, a pregnancy test that produces a positive result if the levels of certain chemical markers are beyond a threshold value,

or a spam filter that classifies messages on the basis of the usage of certain indicative words resulting in a spam score. In these cases, the measurands are the mass of an item, whether a woman is or is not pregnant, and the intention with which a message was sent. The compression of the spring, levels of chemical markers, or spam score are merely part of the measurement procedure or algorithm.

The *true value* $T(A)$ of an item A is the value that should be assigned according to the measurand's definition (JCGM, 2008), and it is a homomorphic map $\mathcal{I} \to \mathcal{S}$ (with $\mathcal{S}$ the measurement scale). It is generally accepted that the true value is more a construct than a concept that can be given operational meaning, and the JCGM (2008) acknowledges that it is not in general a unique value. Because the true value is usually unknowable on principle, a more pragmatic concept is the *reference value* $R(A)$, a value obtained from a higher order measurement system that is accepted by convention to play the role of the true value (JCGM, 2008). In medicine, a similar concept is that of a gold standard, an authoritative test for evaluating diagnostic and screening tests. Reference values and gold standards should be of a higher order of accuracy than the measurement system under study, but it is generally acknowledged in metrology and medicine that they are usually not perfect.

Actual measurements are subject to random measurement error and therefore are a stochastic map $M : \mathcal{I} \times \Omega \to \mathcal{S}$ (with Ω a probability space). *Measurement error* is the discrepancy between a measurement result $M(A)$ and the true value $T(A)$. However, since $T(A)$ has no operational meaning, widely accepted reference values $R(A)$ are often used instead. The numerical expression of measurement error depends on a scale's algebraic structures and operators. Binary and nominal scales are only equipped with the simplest of structures – the equivalence relation – and measurement error therefore takes the form of misclassification: $\{M(A) \neq R(A)\}$. A statistical evaluation is often performed in terms of a probability of misclassification, $P\left[M(A) = a | R(A) \neq a\right]$, which is the probability of classifying something as a, while the reference value is *not* a. Obviously, lower values for these probabilities mean better measurement reliability. For ordinal scales, measurement error can refer to misclassification but, in addition, to whether pairs of items are ordered correctly: $\{M(A) \leq M(B), R(A) \geq R(B)\}$. For measurements on an interval scale, measurement error can be defined as the difference $M(A) - R(A)$ between a measurement result and the reference value, and a statistical evaluation can be in terms of the mean and standard deviation of this difference. For ratio scales, finally, one could even consider the relative measurement error $(M(A) - R(A))/R(A)$.

1.2 Notation and context

Although MSA is of great importance for many fields of science and application, we use the terminology and setting from industry in this thesis. In particular, we consider

quality inspections that give *accept* or *reject* decisions regarding products (or *objects*, *items* or *parts*), in order to reflect whether the *true state* is *conforming* (or *good*) or *nonconforming* (or *defective*) with respect to specifications. We believe that this does not imply a loss of generality and that our findings are relevant outside industry as well.

Measurements are performed by binary measurement systems (also called appraisers). These can be humans, devices, or humans operating devices. The measurement outcome is denoted by Y, and the true state by X. Y can be 0 (reject) or 1 (accept) and X can be 0 (nonconforming) or 1 (conforming). Measurement error is the event $\{Y \neq X\}$, and a correct measurement is the event $\{Y = X\}$.

Just as the manager of the cash center, in industry, we are interested in the misclassification probabilities FAP (false acceptance probability) and FRP (false rejection probability):

$$FAP = P\left[Y = 1 | X = 0\right],$$
$$FRP = P\left[Y = 0 | X = 1\right].$$

In this thesis, we aim to quantify the reliability of binary tests, such as quality inspections in industry. Reliability in the context of MSA means the degree to which a measurement outcome Y truthfully reflects the true value X. The quantification is, in many cases, done by means of the FAP and FRP as defined above.

1.3 Literature review

MSA and metrology are studied in a wide range of disciplines, such as engineering and the natural sciences, medicine and the social sciences. Results are laid down in prescriptive standards such as ISO (1995), AIAG (2003) and JCGM (2008). These standards have a predominant focus on numerical measurements, and binary MSA receives only thin coverage. Even the scientific literature in the field of metrology emphasizes numerical measurement almost exclusively, and attempts to extend frameworks to binary MSA are at an exploratory stage (see, for example, Wilrich, 2010; Suzuki et al., 2013; Bashkansky and Gadrich, 2013).

MSA studies for binary measurements *with* gold standard are widely performed in medicine, where this topic is applied to the study of the accuracy of diagnostic and screening tests; see standard textbooks such as Pepe (2003). Also in industry, binary MSA with gold standard is a reasonably mature science (Danila et al., 2008) and procedures are laid down in industry standards (AIAG, 2003). Approaches for MSA when a gold standard is available are reasonably straightforward, and are based on the idea to have a sample $i = 1, \ldots, I$ of items judged by the system under study

and by the gold standard test, and to compare the resulting $Y_1, \ldots, Y_I$ to the gold standard values $X_1, \ldots, X_I$. The probabilities of false acceptance and rejection *FAP* and *FRP* are estimated from sample proportions (that is, the fractions of good and defective items judged incorrectly). Recent research offers improvements for specific situations, such as when historical data are available (Danila et al., 2013), or when nuisance effects are involved (De Mast et al., 2014).

Binary MSA *without* gold standard, however, is a far less mature science. From the social sciences, there are approaches for binary MSA that do not require a gold standard, such as log-linear modeling (Becker and Agresti, 1992; Hunt et al., 2015) and the kappa statistic (Cohen, 1960; Fleiss, 1971). The latter has been adopted in medicine and engineering as well. However, such approaches evaluate measurements in terms of *agreement*, that is, the event that two repeated judgments Y_{i1} and Y_{i2} of a single item i are identical: $\{Y_{i1} = Y_{i2}\}$. They are not based on the notion of measurement error $\{Y_i \neq X_i\}$, as is the topic of this thesis. Moreover, these approaches have been widely discredited. One of the criticisms is that agreement is driven almost exclusively by the consistency of judgments on conforming products, and therefore fails to reflect whether judgments of defective products are consistent. Also, such approaches require a correction for agreement by chance and a careful sampling strategy, about both of which there is substantial controversy, and results are often characterized as "paradoxical" (Kraemer et al., 2004; De Mast, 2007; Erdmann et al., 2016b).

In both medicine and industry, binary MSA without gold standard is depicted as problematic. See, for instance, Kraemer (1987); Feinstein (2002); Irwig et al. (2002); Knottnerus et al. (2002) in the medical literature and De Mast et al. (2011) in industrial statistics. In recent decades, researchers have attempted to develop approaches for binary MSA that do not require a gold standard. The traditional approach is based on latent class models, which treat the unobservable true state $X_i \in \{0, 1\}$ as a binary latent variable (Hui and Walter, 1980; Boyles, 2001; Van Wieringen and De Mast, 2008; Danila et al., 2008, 2010; Beavers et al., 2011). These models make the assumption of homogeneity within subpopulations of conforming and nonconforming products. That is, every conforming item i has the same probability R_i of being rejected (which is therefore the *FRP*), and likewise, every defective item has the same rejection probability $R_i = P[Y_i = 0]$ (which is therefore equal to $1 - FAP$). This assumption (called a "conditional independence assumption") was found to be unrealistic (Vacek, 1985; Torrance-Rynard and Walter, 1997; De Mast et al., 2011) as, for instance, two items can both be defective, but still have a different *FAP*, when one is more severely defective than the other.

Recently proposed latent variable models allow variation within the subpopulations of conforming and defective items (Qu et al., 1996; Albert et al., 2001). Two types of models can be discerned: random effect models and characteristic curve models. In the random effect models (Danila et al., 2012), the rejection probabil-

ities R_i of conforming items are modeled as random variables with a distribution $F_R^1 = P[R_i \leq r|X = 1]$ on $[0, 1]$, and so are the acceptance probabilities $1 - R_i$ of defective items (with $F_R^0 = P[R_i \leq r|X = 0]$). This leads to variability in the misclassification probabilities, and *FRP* and *FAP* have since been conceived as averages ($FRP = E[R_i|X_i = 1]$ and $FAP = E[1 - R_i|X_i = 0]$). Characteristic curve models (Erdmann, 2012; Erdmann et al., 2016a) assume that variability in the rejection probabilities of conforming and defective items is driven by a continuous, empirical property Z_i. An item's rejection probability is linked to this property by means of a characteristic curve $R_i = q(Z_i)$.

Besides the modeling of variability in the rejection probabilities, another challenge that has acquired attention in the recent literature, is that of extremely unbalanced populations of items. Namely, defect rates in industry are typically very low (as are prevalences of diseases in medicine). As a consequence, a representative sample of items will contain no or very few defective items, which results in problematically large standard errors of the estimates. Artificially increasing the fraction of defective items in the sample results in a nonrepresentative sample, and the resulting estimates for *FAP* and *FRP* are biased substantially (De Mast et al., 2011). Danila et al. (2010, 2012) offer an effective solution to this problem, where items for the MSA study are sampled from the stream of rejected items. Bias in the estimators is avoided by incorporating this initial rejection event as a condition in the likelihood function that is maximized to obtain estimates. This solution is used throughout this thesis.

Albert and Dodd (2004) found that methods for binary MSA without gold standard easily derail when model assumptions are not exactly met. There is no clear indication as to which approach should be preferred in a given situation. Their observation is that different models give substantially different outcomes as a result of the sensitivity to assumption violations. There is no method of detecting these biases, as true values are unobservable.

1.4 Objective and outline of this thesis

Binary measurements and tests are used widely and routinely in medicine, industry and elsewhere. Given the criticality of the consequences of classification errors, reliable methods for MSA are important. The unavailability of a gold standard is a common complication.

There currently is no method that can be used to reliably estimate measurement error from experimental data without gold standard. The unsatisfactory situation is that practitioners have a variety of models they can choose from, and any of them may give substantial and undetectable biases. There are no clear explanations yet for this problem, and it is not yet fully understood.

The focus of this thesis is on the estimation of misclassification probabilities of binary measurement systems, when a gold standard is unavailable. Advances have been made, but there is still no method that estimates misclassification probabilities with consistently proven reliability in the gold standard unavailable case. We believe that not only binary measurements, but also misclassification probabilities, are currently insufficiently understood.

This thesis has three goals. The first is to obtain a better understanding of binary measurement and the quantities *FAP* and *FRP*, and what the unavailability of a gold standard means for the reliability of their estimates. Second, to find alternative quantifiers of measurement error that do give reliable outcomes in the absence of a gold standard. And third, to extend these principles and this method to more general, nominal scales.

Chapter 2 works towards the first goal. It explicitly models the underlying complexities of measurement on a binary scale for a specific real life case. This more fundamental approach provides new insights. It becomes apparent that estimating *FAP* and *FRP* is problematic in this case, but other results still turn out useful for providing practical recommendations to the practitioner.

Chapters 3 and 4 provide a more general proof that the problems that were encountered in Chapter 2 are not specific for that case. Both chapters arrive at the conclusion that a practitioner should not try estimating *FAP* and *FRP* if there is no gold standard. Chapter 3 does so by mathematical analysis, and Chapter 4 shows, with an extensive simulation study, the biases that can occur if one tries to do the estimation anyway. These two chapters aim at the first goal.

Chapters 3 and 4 also show that *FAP* and *FRP* can be decomposed into two parts, in a manner that resembles the conventional decomposition of numerical measurement error into bias and measurement spread. Chapter 5 proposes an adaptive robust method to estimate the binary equivalent of measurement spread only, as is already customary for numerical MSA without gold standard. The method is evaluated in detail and sample size recommendations are provided. Together, Chapters 3, 4 and 5 address the second goal of this thesis.

The third goal is covered in Chapter 6. It is an extension of the method presented in Chapter 5 to measurements on a scale with more than two classes.

1.5 Contributions to scientific literature

Chapter 2 has appeared as an article in *Journal of Quality Technology* (Erdmann et al., 2016a). Dr. Tashi Erdmann brought in the single-appraiser model for this article, supervised the simulations and adviced on data collection for the case study. Execution of the simulation study, the generalization of the single-appraiser model to the multi-appraiser model, programming and estimation were my contribution. Prof.

Jeroen de Mast provided insights in interpretation of outcomes, and both he and Prof. Stefan Steiner greatly improved upon the article with comments and suggestions. Many of the simulations performed for this article were already included in my MSc thesis (Akkerhuis, 2012) and a PhD thesis (Erdmann, 2012). The novelty of this chapter and article is the generalization to the multi-appraiser model, and its application to a real-life problem.

Chapters 3 and 5 were originally two distinct research projects, that have been combined into one article that is submitted to *Journal of Quality Technology* (Akkerhuis et al., 2016a). For Chapter 3, I have introduced the R-domain and showed that recent models only differ with respect to functional form assumptions, which suggests that *FAP* and *FRP* are not identifiable – Prof. Jeroen de Mast put this in a general modeling framework. For the work in Chapter 5, I took the lead and all the chapter's contributions are attributable to me.

Chapter 4 is the empirical underpinning for Chapter 3 and has been submitted to *Measurement* (Akkerhuis et al., 2016b). My contribution was the execution of the simulation study and the design of the infinite sample technique. Prof. Jeroen de Mast and Dr. Erdmann have supervised the research design and interpretation of results and helped in combining the large amount of findings into a smaller set of general conclusions.

The contributions in the final chapter, Chapter 6, are attributable to me, although its present form was reached by valuable comments by Prof. Jeroen de Mast. It will appear in *Quality and Reliability Engineering International* (Akkerhuis and De Mast, in press).

Section 1 in this introduction is based on the theoretical framework in De Mast et al. (2014), an article in *Quality Engineering* that I co-authored.

Chapter 2

A mechanistic model for binary measurements

One of the goals of this thesis is to obtain a better understanding of binary measurement and the false acceptance and rejection probabilities (FAP and FRP), and what the unavailability of a gold standard implies for the reliability of their estimates (see Chapter 1). The case study in this chapter contributes to the understanding, by explicitly modeling the mechanics underlying a particular binary test. Estimation of FAP and FRP proves to be problematic, which is an important implication of the unavailability of a gold standard. Later chapters will show that this implication does not only hold in this particular case.

2.1 Introduction

This chapter reports on a case study from an optical inspection system in a plant for car parts. The subjects of testing are sensors that are applied in cars, and in particular, the soldering connection between two of the sensors' components. The inspection system decides to accept ($Y = 1$) or reject ($Y = 0$) a sensor, based on the quality of the soldering connection. The soldering quality is based on whether the misalignment Z of the two components is above or below an upper specification limit USL. A sensor conforms to specifications ($X = 1$) if $Z \leq USL$, and it is considered defective ($X = 0$) if $Z > USL$. The reliability of these inspections can be quantified by the false acceptance and false rejection probabilities, which can be expressed as a function of the misalignment Z:

$$FAP = P[Y = 1|X = 0] = P[Y = 1|Z > USL],$$
$$FRP = P[Y = 0|X = 1] = P[Y = 0|Z \leq USL].$$

In this chapter, we deal with a typical combination of three challenges. The first challenge is that the true conditions of the items (X and Z) are practically unobservable. This rules out the use of traditional methods for estimating the FAP and FRP (as described in AIAG, 2003; Pepe, 2003; Danila et al., 2008), which assume the availability of a so-called gold standard: a higher order, authoritative measurement that is accepted to constitute a faithful representation of the measurand.

For situations where no gold standard is available, the literature proposes methods based on latent class models (Hui and Walter, 1980; Boyles, 2001; Van Wieringen and De Mast, 2008; Danila et al., 2010 and many more). Such methods are essentially based on the premise that the FAP and FRP are constant across items in the populations of defective and good items, respectively. Technically, such methods assume that the inspections Y are independent and identically distributed conditional on whether the items are defective or good. However, this assumption is often too simplistic: De Mast et al. (2011) show that a true value related to a continuous property, such as misalignment in our example, violates such conditional independence assumptions, which leads to biased estimators (possibly substantially so). Thus, the second challenge is in avoiding the assumption that the FAP and FRP are constants in the populations of defective and good items.

The third challenging aspect of the situation that we consider is that the prevalence of defects in industrial processes is typically very low. Consequently, a sample from the population of all produced items would contain no or only very few defective items, resulting in impractically large standard errors in the estimated FAP.

This combination of three challenges is quite common, and this chapter proposes an approach based on latent trait models. A similar model was introduced for ordinal classifications in De Mast and Van Wieringen (2010) and is related to item-response theory models (Lord, 1980). In the latent trait model, the probability of rejection is not assumed constant within the populations of defective and good items. Instead, it is a function of an unobserved continuous property Z, and the dependence is modeled by means of a characteristic curve $q(z) = P[Y = 0|Z = z]$. This curve is estimated from data collected in an MSA experiment. FAP and FRP or comparable metrics can be determined from these two functions, as shown later. Due to the third challenge (defects are extremely rare in the population of items), a random sample of items will contain no or only very few defective items. Estimation of the relevant part of the characteristic curve q, however, requires a sample with fair numbers of good and defective items. Our solution involves combining samples from various origins (the total items population, the stream of rejected items, and historical data about the

rejection rate) and incorporating the sampling origin in the likelihood function to correct for a potential bias due to nonrandom sampling.

The purpose of this chapter is to elaborate this approach for binary inspections and to explore to what extent it provides an effective solution to the combination of the three mentioned challenges. We elaborate our approach in Section 2.2. Section 2.3 presents an evaluation of the approach on the basis of simulation and asymptotics. Section 2.4 describes the motivating case about optical inspections of car parts, and we draw conclusions in the final section.

2.2 Methodology

2.2.1 Statistical model

In this subsection we present our model. We consider MSA experiments in which a sample of items $i = 1, ..., I$ is (repeatedly) appraised by each of one or more appraisers $j = 1, ..., J$. Repeated appraisals of item i by appraiser j are indexed by $k = 1, ..., K_{ij}$. The subscript on K_{ij} indicates that the number of repetitions may differ over items and appraisers, and $K_{ij} = 0$ means that appraiser j did not measure item i. We denote the result of the kth appraisal by appraiser j for item i as Y_{ijk}. Thus, the outcome of the MSA experiment is the array of zeros and ones given by $\boldsymbol{Y} = \{Y_{ijk}\}_{i=1,...,I;j=1,...,J;k=1,...,K_{ij}}$.

The binary appraisals under study aim to reflect an unobservable, continuous property Z. Without loss of generality, we assume that we only have an upper specification limit (USL). An item i is "good" if $Z_i \leq USL$, and the intended inspection outcome in that case is $Y_{ijk} = 1$ ("accept"). If $Z_i > USL$, the item is "defective" ($X = 0$) and the intended outcome is $Y_{ijk} = 0$ ("reject"). The Z_i are assumed to be independently distributed and, in the population of produced items, have a normal distribution. The location and scale of the latent Z-continuum are arbitrary and, without loss of generality, we set $\mathbb{E}[Z_i] = 0$ and $\mathrm{Var}[Z_i] = 1$. Without such restrictions, the latent variable model is unidentifiable. Thus, the density of Z_i in the population of items is the standard normal denoted ϕ.

We assume that besides Z_i, there are no other properties of the items and no environmental factors that induce dependencies among repeated appraisals. In particular, we assume that, conditional on Z_i, the $Y_{ij1}, ..., Y_{ijK_{ij}}$ are independent and identically Bernoulli distributed. Careful experimental design may enable this assumption to be fulfilled, for example, by experimental randomization and, in the case of human appraisers, by presenting the items in a way they cannot be recognized from the previous inspection. See Chapter 4 or De Mast et al. (2014) for a discussion about handling nuisance variables.

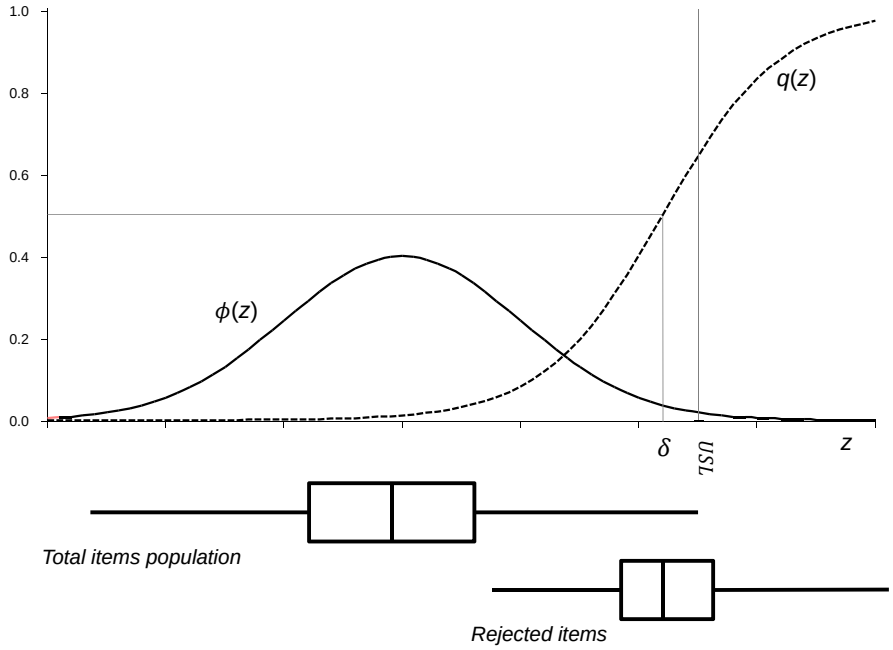


Figure 2.1: Probability density $\phi(z)$ and characteristic curve $q_j(z)$. Boxplots represent the distribution of Z in a representative sample from the total items population and in a sample from the rejected items.

We define $q_j(z) = P[Y_{ijk} = 0 | Z_i = z]$ as the characteristic curve of appraiser j. Our initial choice for $q_j(x)$ is the logistic function

$$q_j(z) = \frac{1}{1 + \exp\{-\alpha_j(z - \delta_j)\}}, \alpha_j > 0. \tag{2.1}$$

The curve's inflection point δ_j, which is also the point where $q_j(z) = 1/2$, can be interpreted as the decision threshold that appraiser j appears to apply: items with $Z_i > \delta_j$ are more likely to be rejected than accepted. The parameter $\alpha_j > 0$ is a discrimination parameter for appraiser j, determining the steepness of the curve. Larger values of α_j correspond to a steeper curve and better reliability of the inspection results. The characteristic curve q_j is symmetric about δ_j. Figure 2.1 illustrates the model. The misclassification probabilities are not equal for all items, but depend on the misalignment Z. The average probabilities (weighted by the density of Z_i in the

population of items) for appraiser j are:

$$\begin{aligned} FAP_j &= \frac{\int_{USL}^{\infty} (1 - q_j(z))\,\phi(z)\,\mathrm{d}z}{\int_{USL}^{\infty} \phi(z)\,\mathrm{d}z}, \\ FRP_j &= \frac{\int_{-\infty}^{USL} q_j(z)\,\phi(z)\,\mathrm{d}x}{\int_{-\infty}^{USL} \phi(z)\,\mathrm{d}z}. \end{aligned} \tag{2.2}$$

In the situation we consider in this chapter, the misalignment Z_i is unobservable and, as a consequence, the USL is ill defined. Consequently, also FAP_j and FRP_j are ill defined. We propose to work instead with alternatives proposed by De Mast and Van Wieringen (2010). The inconsistent acceptance probability (IAP_j) and inconsistent rejection probability (IRP_j) are the probabilities that appraiser j's classification is inconsistent with his or her own decision threshold δ_j.

$$\begin{aligned} IAP_j &= \frac{\int_{\delta}^{\infty} (1 - q_j(z))\,\phi(z)\,\mathrm{d}z}{\int_{\delta}^{\infty} \phi(z)\,\mathrm{d}z}, \\ IRP_j &= \frac{\int_{-\infty}^{\delta} q_j(z)\,\phi(z)\,\mathrm{d}z}{\int_{-\infty}^{\delta} \phi(z)\,\mathrm{d}z}. \end{aligned} \tag{2.3}$$

Whereas FAP_j and FRP_j express both the systematic component of classification error (that is, $|\delta_j - USL|$) and the random component (the degree to which classifications randomly deviate from an appraiser's own δ_j), IAP_j and IRP_j express the random component only. This can be seen from the following decomposition of FRP_j, where we assume that $\delta_j \leq USL$ (and similar decompositions can be given for FAP_j and for $\delta_j > USL$):

$$\begin{aligned} FRP_j = P[Y = 0 | Z \leq \delta_j]\, P[Z \leq \delta_j | Z \leq USL] + \\ P[Y = 0 | \delta_j < Z \leq USL]\, P[\delta_j < Z \leq USL | Z \leq USL]. \end{aligned}$$

The last term is the contribution to FRP_j due to systematic classification error, which is determined by the distance between δ_j and USL. The first term is IRP_j (both terms

are multiplied by probability weights). Without gold standard measurement, IAP_j and IRP_j, rather than FAP_j and FRP_j, are often all that one could hope to estimate (see Chapter 3).

The relevant aspects of measurement reliability can now be framed in technical terms. Random measurement error can be quantified by IAP_j and IRP_j. Systematic differences between appraisers can be quantified by differences in their decision thresholds δ_j. Further, the percentage of defective items reaching the customer is $P[Z > USL|Y = 1]$ and the percentage of falsely rejected ("conforming") items in the stream of rejects is $P[Z \leq USL|Y = 0]$. If USL is known (or assumed to equal the δ_j of an appraiser j who is taken as reference standard), these probabilities can be determined from:

$$P[Z > USL|Y = 1] = \frac{\int_{USL}^{\infty} (1 - q_j(z))\,\phi(z)\,\mathrm{d}z}{\int_{-\infty}^{\infty} (1 - q_j(z))\,\phi(z)\,\mathrm{d}z},$$

$$P[Z \leq USL|Y = 0] = \frac{\int_{-\infty}^{USL} q_j(z)\,\phi(z)\,\mathrm{d}z}{\int_{-\infty}^{\infty} q_j(z)\,\phi(z)\,\mathrm{d}z}.$$

2.2.2 Sampling strategy and estimation

We propose maximum likelihood estimators for the parameters $(\alpha_j, \delta_j), j = 1, ..., J$ of the characteristic curves q_j (see 2.1). First we describe the sampling strategy that we propose.

A representative sample from the total population of items is the obvious choice, but on second thought this turns out to be problematic. As mentioned in the introduction, rejection rates in typical manufacturing processes are low and the steep part of q_j (around its inflection point δ_j) will usually be in the remote tail of the density ϕ of Z in the items population. Consequently, a representative sample from the total items population of reasonable size will contain no or only a few Z-values close to or to the right of δ_j. This results in very large standard errors in the estimated IAP_j. Figure 2.1 illustrates the point: the first boxplot below the graph shows the quartiles of the probability density of Z for items in such a sample (the box represents the 25%, 50% and 75% quartiles and the whiskers delineate a 99% interval). It can be seen that only a small fraction or even none of the items have values in the steep part of $q_j(z)$.

To estimate the characteristic curves close to and to the right of δ_j with acceptably small standard errors, one needs a sample with more evenly spread z-values. We

propose to take samples from various sampling sources and combine them to have a more balanced sample. Besides the total population of items (Tot-sample), we also propose sampling from the population of rejected items (Rej-sample) and incorporating aggregate information from historical data (His-sample). It is essential that, during the period over which the items are sampled, the circumstances are constant. This may be realized by ensuring the three subsamples are taken over the same time period.

The idea to sample items from the stream of rejects was proposed by Danila et al. (2010). They called the resulting data a conditional sample, as the sampling distribution of items sampled from the population of rejected items is obtained by conditioning on the initial rejection decision. This approach typically leads to a sample that has a larger proportion of items in the steep part of the characteristic curve (see the boxplot labeled "Rejected items" in Figure 2.1). For an item i sampled from the stream of rejected items, let $d_i \in \{1, ..., J\}$ be the appraiser who rejected it, and let $Y_{id_i0} = 0$ denote the event of this rejection. Given that item i has been rejected by appraiser d_i, the sampling distribution of Z_i is

$$F^{d_i}(z) = P[Z_i \leq z | Y_{id_i0} = 0] = \frac{\int_{-\infty}^{z} q_{d_i}(t)\,\phi(t)\,\mathrm{d}t}{\int_{-\infty}^{\infty} q_{d_i}(t)\,\phi(t)\,\mathrm{d}t} \tag{2.4}$$

with probability density f^{d_i}.

Besides the Tot- and Rej-samples, a third data source is a historical dataset of (single) inspection results ("reject" or "accept") of a large number of items, typically summarized as a total count and a number of rejected items. Danila et al. (2010, 2012) called this baseline data, and showed that it substantially increases the precision of the estimators in the latent class models they discussed. A historical rejection rate is typically easy to obtain. If the inspection system has been in use for a while, such data are typically available, and otherwise, they can be obtained during the collection of the Rej-sample for the MSA study (which typically involves several thousand inspections before a sufficient number of rejected items are obtained). Conditional on Z_i, the likelihood of all outcomes for item i is

$$P[Y_{i11} = y_{iJ1}, ..., Y_{iJK_{iJ}} = y_{iJK_{iJ}} | Z_i = z_i] = \prod_{j=1}^{J} \prod_{k=1}^{K_{ij}} q_j(z_i)^{1-y_{ijk}} (1 - q_j(z_i))^{y_{ijk}}.$$

This expression uses conditional independence of repeated appraisals. To obtain the unconditional probability we integrate out the latent variable Z_i weighted by its probability density. For items in the Tot-sample and His-sample this is $f^0 = \phi$. For items

in the Rej-sample this is $f^{d_i}(z)$ as in (2.4). This gives

$$P\left[Y_{i11}=y_{iJ1},...,Y_{iJK_iJ}=y_{iJK_iJ}\right]= \int_{-\infty}^{\infty} f^{d_i}(z)\prod_{j=1}^{J}\prod_{k=1}^{K_{ij}} q_j(z)^{1-y_{ijk}}\left(1-q_j(z)\right)^{y_{ijk}}\mathrm{d}z,$$

where $d_i=0$ for items in the Tot- and His-samples, and $K_{ij}=1$ if i is in the His-sample. The resulting loglikelihood, with parameter vector $\boldsymbol{\theta}=\{(\alpha_j,\delta_j)\}_{j=1,...,J}$ is

$$L(\boldsymbol{\theta}|\boldsymbol{y})=\sum_{i=1}^{I}\log\left[\int_{-\infty}^{\infty} f^{d_i}(z)\prod_{j=1}^{J}\prod_{k=1}^{K_{ij}} q_j(z)^{1-y_{ijk}}\left(1-q_j(z)\right)^{y_{ijk}}\mathrm{d}z\right],$$

where $\boldsymbol{y}=\{y_{ijk}\}_{i=1,...,I;j=1,...,J,k=1,...,K_{ij}}$ represents the data. For concise representations of the experimental outcomes $\boldsymbol{Y}$ and efficient calculation of the loglikelihood, we use an equivalent expression in terms of response pattern frequencies.

A response pattern $\boldsymbol{R_i}=\left(\sum_{k=1}^{K_{ij}} Y_{ijk},\sum_{k=1}^{K_{ij}}\left(1-Y_{ijk}\right)\right)_{j=1,...,J}$ is a $J\times 2$ matrix, in which the elements $R_i[j,1]$ are the number of times item i is accepted by appraiser j, and $R_i[j,2]$ is the number of times it is rejected by appraiser j, and $R_i[j,1]+R_i[j,2]=K_{ij}$. As before, let $d_i=0$ if item i is in the Tot- or His-sample, and $d_i\in\{1,...,J\}$ be the appraiser who initially rejected the item if i is in the Rej-sample. The response pattern frequencies are $e(\boldsymbol{r},d)=\{\#i|\boldsymbol{R}_i=\boldsymbol{r},d_i=d\}$ with $d\in\{0,...,J\}$ and $\boldsymbol{r}$ an $J\times 2$ matrix with elements in $\mathbb{N}$. By tabulating $e(\boldsymbol{r},d)$, the data can be displayed concisely, and the loglikelihood can be rewritten to contain fewer integrals:

$$L(\theta|\boldsymbol{y})=\sum_{d=0}^{J}\sum_{\boldsymbol{r}\in\mathbb{N}^{A\times 2}} e(\boldsymbol{r},d)\times\log\left[\int_{-\infty}^{\infty} f^{d}(z)\prod_{j=1}^{J} q_j(z)^{r[a,2]}\left(1-q_j(z)\right)^{r[a,1]}\mathrm{d}z\right]. \tag{2.5}$$

We maximize the loglikelihood with the interior point algorithm (Mehrotra, 1992), as implemented in the function "FindMaximum" in software package Wolfram Mathematica 8 (2010). We find starting values using the Nelder-Mead algorithm (Nelder and Mead, 1965), as implemented in "NMaximize" in Wolfram Mathematica 8. The integrals are approximated numerically using adaptive quadrature (Rice, 1975) as implemented in "NIntegrate" in Wolfram Mathematica 8. Once the parameters α_j and δ_j have been estimated, they can be plugged into $q_j(x)$ in (2.3) to obtain the estimates for IAP_j and IRP_j.

Note that the convergence time rapidly increases in the number of appraisers J. The increasing number of parameters and exponentially increasing number of

response patterns in J leads to a large number of integrals to evaluate. We find that for $J > 1$, the optimization may take a half hour, and even more for some choices of the functional form for $q_j(x)$.

Standard errors of the maximum likelihood estimators $\widehat{\boldsymbol{\theta}}$ and of $\widehat{IAP}_j$ and $\widehat{IRP}_j$ can be obtained by Monte Carlo simulation, but in view of the above this becomes practically undoable if $J > 1$. We recommend approximating the covariance matrix $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\theta}}}$ of $\widehat{\boldsymbol{\theta}}$ on the basis of the observed Fisher information matrix:

$$\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\theta}}} = \left(-\frac{\partial^2}{\partial \boldsymbol{\theta}^2} L(\boldsymbol{\theta}|\boldsymbol{y}) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right)^{-1}. \tag{2.6}$$

The covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\widehat{IAP}_j, \widehat{IRP}_j}$ of $\widehat{IAP}_j$ and $\widehat{IRP}_j$ is then approximated by the Delta method

$$\left(-\frac{\partial^2}{\partial \boldsymbol{\theta}^2} (IAP_j, IRP_j) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right) \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\theta}}} \left(-\frac{\partial^2}{\partial \boldsymbol{\theta}^2} (IAP_j, IRP_j) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right)^T, \tag{2.7}$$

which uses a linear approximation to the functions $IAP_j(\boldsymbol{\theta}), IRP_j(\boldsymbol{\theta})$.

2.2.3 Model diagnostics

For assessment of the fit of the model we can use standard techniques from the latent variable modeling literature. We briefly mention them here, and demonstrate their use in the car parts case described later on. For residual analysis, one may compare the observed frequencies $e(\boldsymbol{r}, d)$ of the response patterns to the frequencies $\eta(\boldsymbol{r}, d)$ predicted by the fitted model, either as raw differences or as Freeman-Tukey variance-stabilized residuals (Formann, 2003). To test for lack of fit, one can apply a likelihood ratio test (which is known as the G-test when response pattern frequencies are concerned). This test is preferred over the Pearson χ^2 test when one or more expected cell counts are below 5 (Kallenberg et al., 1985). The test statistic is based on the likelihood ratio:

$$G = 2 \sum_{d=0}^{J} \sum_{\boldsymbol{r} \in \mathbb{N}^{J \times 2}} e(\boldsymbol{r}, d) \log \frac{e(\boldsymbol{r}, d)}{\eta(\boldsymbol{r}, d)} \sim \chi^2(df).$$

The number df of degrees of freedom of the chi-square distribution is the difference between the number of parameters needed for a fully saturated model and the number of parameters in the model ($2J$ for logistic curves). The calculation is straightforward but confusing to define for the general case; we will instead demonstrate the calculation for the case. By simulation, we have evaluated the test's power in detecting one form of a lack of fit in particular, namely, that q is not symmetrical about δ, but has

different slopes to the left and right of δ (details can be found in Akkerhuis, 2012). The test turns out to have useful power for detecting departures from symmetry. If asymmetry is detected, one could fit a loglogistic function instead:

$$q(z) = \frac{1}{1 + (\alpha(z - \mu))^{\beta}} \mathbf{1}_{z \geq \mu}, \tag{2.8}$$

where the indicator function $\mathbf{1}_{z \geq \mu}$ indicates that the curve is zero for $z < \mu$. This family of curves is more flexible and can adapt to varying degrees of asymmetry. We demonstrate this later on.

2.3 Suitable sample sizes and robustness

We conducted a number of simulations to assess the properties of the estimators proposed in the previous section. The aims of these studies were to establish guidelines for choosing sample sizes and to assess the robustness of the estimators against model misspecification. The studies are limited to the case of a single appraiser ($J = 1$). Finite-sample properties were established by Monte Carlo simulation. In addition, we derived the asymptotic distribution of the maximum likelihood estimators, by replacing the data in (2.6) by their expected values. For $I \to \infty$ we have that $\widehat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \left(-E_{\boldsymbol{Y}} \left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} L(\boldsymbol{\theta} | \boldsymbol{Y}) \right] \right)^{-1}. \tag{2.9}$$

For the interested reader, a detailed report can be found in Erdmann (2012) or Akkerhuis (2012). In this thesis, we briefly summarize the results.

2.3.1 Optimal proportions of sample sources

The total sample consists of $I = I^{\text{Tot}} + I^{\text{Rej}} + I^{\text{His}}$ items, with I^{Tot}, I^{Rej}, and I^{His} the sizes of the Tot-, Rej- and His-samples. First, we investigated which proportion of sample sources is optimal. As noted before, a large historical dataset of rejections is typically available or can be obtained while collecting the Rej-sample. For this reason, we assumed a large His-sample size of $I^{\text{His}} = 100,000$. To see how much the His-sample contributes we also investigated the case that $I^{\text{His}} = 0$.

The study demonstrated that the most precise estimates of *IAP* and *IRP* are obtained when the sample consists solely of a Rej-sample (that is, $I^{\text{Tot}} = 0$) supplemented with a large His-sample. Further details of the study are as follows. We considered sample sizes $I^{\text{Tot}} \in \{0, 10, ..., 200\}$ and $I^{\text{Rej}} = 200 - I^{\text{Tot}}$. In both the data-generating process and the estimated model, a standard normal distribution was used

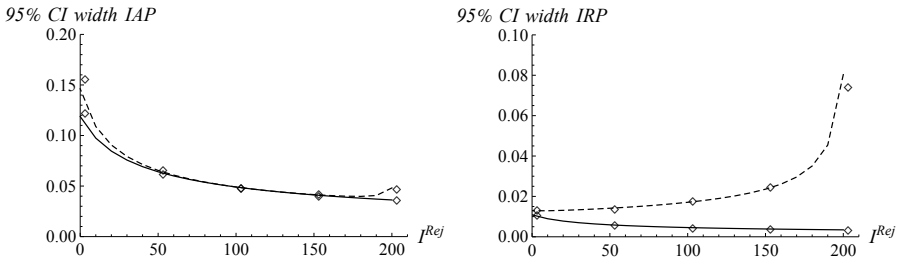


Figure 2.2: 95% confidence interval widths for an inspection system with $\alpha = 5, \delta = 2, IAP = 0.2154, IRP = 0.0125$ for various I^{Rej}, I^{His}, $\left(I^{\text{Tot}} = 200 - I^{\text{Rej}}\right)$. Lines: results obtained through the asymptotic distribution of the MLE (dashed: $I^{\text{His}} = 0$; solid: $I^{\text{His}} = 10^5$). Diamonds: results obtained by simulation.

for Z and a logistic curve with $\alpha = 5, \delta = 2$ for q. These choices imply true probabilities of $IAP = 0.2154$ and $IRP = 0.0125$. Note that the IAP is (much) larger than the IRP, which is typical when δ is in the far tail of the distribution of Z. Conditional on $Z > \delta$, $\phi(z)$ has the most probability mass close to δ (which corresponds to the range of hard-to-judge items), while conditioning on $Z \leq \delta$, most probability mass is around 0 (with easy-to-judge items).

The precision of the estimators in each scenario was determined from their asymptotic covariance matrix (2.9) and also from Monte Carlo simulation (2500 runs per scenario). Precision of the estimators was quantified as the width of empirical 95% confidence intervals for IAP and IRP (based on the 2.5 and 97.5 percentiles of the 2500 realizations of $\widehat{IAP}$ and $\widehat{IRP}$).

Figure 2.2 gives plots of the 95% confidence interval width for $\widehat{IAP}$ and $\widehat{IRP}$ as a function of I^{Rej} with and without a His-sample. The lines in the figure are based on asymptotic standard errors and the diamonds on Monte Carlo simulation. The conclusions were corroborated for other values of α and δ (based on a study with a more limited range of values for I^{Tot}, I^{Rej} and I^{His}).

2.3.2 Precision as a function of I^{Rej} and K

Following the recommendations above, the MSA experiment should involve a sample of I^{Rej} items from the stream of rejects, which are appraised K times, and a historical rejection rate estimated from I^{His} appraisals. To choose an appropriate sample size I^{Rej} and number of repeated appraisals K, we determine their effect on the precision of the estimators.

As before, we set $I^{\text{His}} = 100,000$. By Monte Carlo simulation, we investigated 7×7 combinations in the ranges $I^{\text{Rej}} \in \{50, 75, \ldots, 200\}$ and $K \in \{3, 5, \ldots, 15\}$. In

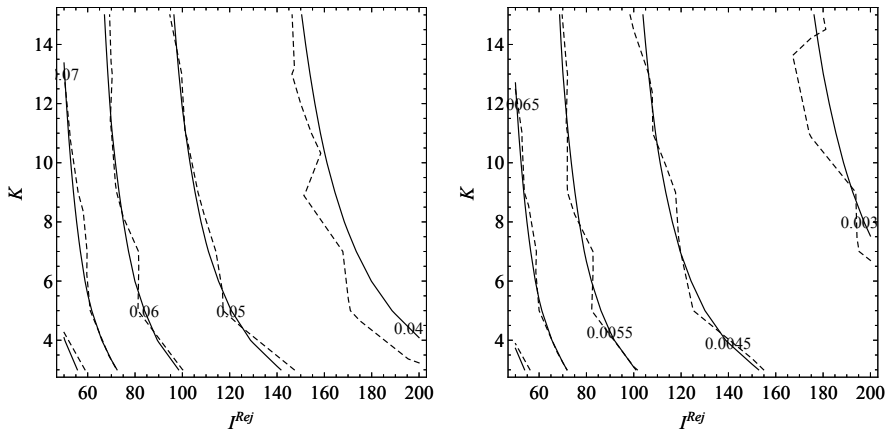


Figure 2.3: Level plots of 95% empirical confidence interval widths for *IAP* and *IRP* estimates, for an inspection system with $\alpha = 5, \delta = 2, IAP = 0.2154, IRP = 0.0125$. Solid lines: results obtained through the asymptotic distribution of the MLE. Dashed lines: results obtained by simulation.

addition, we derived asymptotic results for 16×13 combinations in the same ranges. We obtained results for all four combinations of $\alpha = 5, 12$ and $\delta = 2, 3$, which we think represent typical binary inspection systems in industry in terms of reliability and rejection rate. Figure 2.3 gives level plots of the empirical 95% confidence interval width as a function of I^{Rej} and K for $\alpha = 5$ and $\delta = 2$. Level plots of the precision for other (α, δ) combinations gave comparable results.

The results obtained by simulation (dashed lines) are close to the asymptotic results (solid lines). We conclude that the asymptotic properties provide a useful approximation to the finite-sample properties of the estimators. This is an important result for the calculation of standard errors for the estimators.

The marginal effects of I^{Rej} and K are diminishing. In particular, for these values of α, δ, it seems inefficient to do more than $K = 7$ repeated appraisals. The required sample size depends on the desired precision, and plots such as in Figure 2.3 can be used as a reference. In general, we recommend $I^{\text{Rej}} \geq 150$ because, in the considered ranges, this ensures a confidence interval width of at most 50% of the *IAP* or *IRP* (assuming $K = 7$), which seems reasonable to us.

2.3.3 Robustness to misspecification

Finally, we report on the estimation procedure's robustness against model misspecification. As mentioned before, the G-test is powerful in detecting asymmetry in $q(z)$,

and in the case we will demonstrate how to handle such situations with loglogistic characteristic curves. Therefore, we focus on misspecification of the distribution of Z here. We simulated 20 scenarios where we drew realizations of Z $\left(I^{\text{Rej}} = 200, I^{\text{His}} = 100000\right)$ from skewed and leptokurtic distributions instead of the normal (and applied a Bernoulli trial to simulate drawing from the stream of rejects for the Rej realizations). For each realization of Z, we drew $K = 9$ realizations of Y by applying a logistic characteristic curve (with parameters α, δ). We fitted our model (with logistic curve for q and a standard normal distribution for Z) to these data, and compared the estimated $\widehat{IAP}$ and $\widehat{IRP}$ to the true values.

Each scenario in Table 2.1 is determined by a distribution for Z (standard normal, standard lognormal, $\chi^2(1)$, $t(3)$, or $t(7)$) and values for α and δ. These choices imply a corresponding IAP, IRP and reject probability $p_0 = P[Y = 0]$. We could not choose the scenarios such that they have the same IAP, IRP and reject probability across distributions. Namely, given a distribution, there are only two free parameters (α and δ), and in addition, some ratios of IAP and IRP are not possible for some distributions. Instead, rows in Table 2.1 were chosen to roughly represent the situations of poor and good measurement reliability and medium and low reject rates. Poor and good measurement reliability are defined here in terms of the 99% part of the characteristic curve (the "grey area" where inspection results are uncertain) compared with the 99% part of the distribution of Z:

$$\frac{q^{-1}(0.995) - q^{-1}(0.005)}{F^{-1}(0.995) - F^{-1}(0.005)} = \rho.$$

We took $\rho = 0.30$ for poor reliability and $\rho = 0.15$ for good reliability, and $p_0 = 0.005$ for a low reject rate, and $p_0 = 0.010$ for a medium reject rate. Table 2.1 shows the α and δ values that give these results for each of the five distributions given in the columns, and also the implied IAP and IRP. We performed 1000 simulation runs for each scenario. The table gives the absolute deviation of the average $\widehat{IAP}$ and $\widehat{IRP}$ from their true values, and an asterisk indicates that a bias is significantly different from 0 (based on a t-test).

We have also investigated robustness against multidimensional (multivariate normal and mixtures) distributions by Monte Carlo simulation, and robustness against skew-normal and t-distributions based on the asymptotic distribution of the covariance estimators (not included in this thesis).

For all forms of misspecification that we investigated, biases in the estimates remain below 0.0100 for IAP, and 0.0007 for IRP (except for one case of extreme negative skewness; in this case, the absolute bias of $\widehat{IAP}$ is larger, but relative to the true IAP, the bias is still modest). In all cases considered, biases remain small compared to the true IAP and IRP. In conclusion, the estimation procedure seems reasonably robust to misspecification of the distribution of Z.

	$Z \sim N(0,1)$		$Z \sim t(3)$		$Z \sim t(7)$		$Z \sim \log N(0,1)$		$Z \sim \chi^2(1)$	
	IAP	*IRP*	*IAP*	*IRP*	*IAP*	*IRP*	*IAP*	*IRP*	*IAP*	*IRP*
$\rho = 0.15$,	$\alpha = 13.7$,		$\alpha = 6.04$,		$\alpha = 10.1$,		$\alpha = 5.40$,		$\alpha = 8.96$	
$p_0 = 0.0005$	$\delta = 2.60$		$\delta = 5.87$		$\delta = 3.52$		$\delta = 13.2$		$\delta = 7.89$	
True value	0.1194	0.0009	0.0489	0.0003	0.0789	0.0005	0.0266	0.0002	0.0398	0.0002
Bias	0.0003	0.0000	0.0013*	0.0000*	0.0014*	0.0000*	0.0080*	0.0001*	0.0010*	0.0000
$\rho = 0.15$,	$\alpha = 13.7$,		$\alpha = 6.04$,		$\alpha = 10.1$,		$\alpha = 5.40$,		$\alpha = 8.96$,	
$p_0 = 0.0005$	$\delta = 2.35$		$\delta = 5.48$		$\delta = 3.02$		$\delta = 10.3$		$\delta = 6.65$	
True value	0.1120	0.0016	0.0589	0.0008	0.0824	0.0012	0.0312	0.0004	0.0404	0.0005
Bias	0.0006	0.0000	0.0010*	0.0001*	0.0009*	0.0001*	0.0011*	0.0000*	0.0004	0.0000
$\rho = 0.15$,	$\alpha = 6.85$,		$\alpha = 3.02$,		$\alpha = 5.04$,		$\alpha = 2.70$,		$\alpha = 4.48$	
$p_0 = 0.0005$	$\delta = 2.67$		$\delta = 5.96$		$\delta = 3.60$		$\delta = 13.2$		$\delta = 7.93$	
True value	0.2012	0.0019	0.0875	0.0007	0.1366	0.0012	0.0504	0.0003	0.0744	0.0005
Bias	0.0001	0.0000	0.0041*	0.0001*	0.0062*	0.0001*	0.0006*	0.0000	0.0006*	0.0000*
$\rho = 0.15$,	$\alpha = 6.85$,		$\alpha = 3.02$,		$\alpha = 5.04$,		$\alpha = 2.70$,		$\alpha = 4.48$	
$p_0 = 0.0005$	$\delta = 2.41$		$\delta = 4.69$		$\delta = 3.10$		$\delta = 10.3$		$\delta = 6.68$	
True value	0.1912	0.0035	0.1030	0.0018	0.1422	0.0026	0.0583	0.0008	0.0753	0.0010
Bias	0.0001	0.0000	0.0070*	0.0002*	0.0065*	0.0002*	0.0007*	0.0001	0.0014*	0.0000*

Table 2.1: Bias (absolute value) of the estimators of *IAP* and *IRP* (based on $I^{\text{Rej}} = 200$; $K = 9$; $I^{\text{His}} = 100,000$) in 20 scenarios. *Estimated bias significantly different from zero.

2.4 MSA Study at a car parts manufacturer

2.4.1 Background of the MSA study

We apply the proposed approach in an MSA experiment conducted to evaluate an optical inspection system at a car parts manufacturer. The parts in question contain an integrated circuit, about 1 cm in diameter, on which several electronic components are connected and soldered. After assembly, the parts are inspected for soldering errors. This check is performed by a team of operators, but a visual inspection machine called Automated Optical Inspection (AOI) has been purchased to replace the operators' inspections in the future. Currently, the AOI is operational, but the operators still perform the inspections. This situation will continue until the AOI is deemed to function satisfactorily.

The AOI inspects a variety of properties of a soldered part, and produces a binary decision in terms of "accept" or "reject" and, in the latter case, a specification of one or more failure modes. Our evaluation focused on the reliability of decisions about one failure mode in particular, namely, misalignment due to soldering faults. Here, the relevant property is the misalignment Z of a clip and the pad to which it is to be fastened. Due to the very small scale and uneven three-dimensional shape of the involved components, misalignment is very hard to measure directly. In addition, there is considerable ambiguity as to its precise definition, and there is no clearly defined upper specification limit (*USL*) that demarcates the acceptable range. The AOI's evaluation is based on a digital photo of the part, which is then analyzed by a

proprietary algorithm.

The judgments by the team of operators were accepted by the company as the de facto standard, and are used to determine whether the AOI's decisions are correct. However, since we found that there was occasional disagreement among the operators themselves, we could not treat them as a gold standard. Rather than fitting a characteristic curve for each operator individually, we treated the team of operators as a single appraiser and appraisals by individual team members as repetitions. Thus, the study involved $J = 2$ appraisers: the AOI ($j = \text{AOI}$) and the team of operators ($j = \text{opr}$). The purpose of the reliability study was to evaluate:

- Variability of the AOI's decisions, represented by the probabilities IAP_{AOI} and IRP_{AOI};
- Variability of the operators decisions, represented by IAP_{opr} and IRP_{opr}; and
- The systematic difference, if any, between the decisions of the AOI and the operators, represented by $\delta_{\text{AOI}} - \delta_{\text{opr}}$.

Due to a combination of changing objectives, advancing insight, practical limitations and communication problems, the sample and dataset that we obtained are not optimal, but they do allow a useful analysis. Based on the results of the simulation studies reported in Sections 2.3.1 and 2.3.2, a sample of $I^{\text{Rej}} = 150$ parts that had been rejected by the AOI (the "Rej-sample") was collected. The parts in this sample were inspected $K_{\text{AOI}} = 7$ times by the AOI, and once by each of $K_{\text{opr}} = 3$ operators. The AOI can measure seven parts simultaneously in different slots, and the parts were randomized over these slots during repeated measurements. A historical rejection rate of 0.050% for the AOI was also available (based on 1271 rejections out of $I^{\text{His}} = 254,200$ inspected parts). Moreover, $I^{\text{Tot}} = 100$ parts from the total parts population (the "Tot-sample") were included, which were measured 7 times by the AOI, but not by the operators. The inclusion of these 100 parts cannot be motivated from the results reported in the previous sections. At the time when the MSA experiment was designed, we believed this additional subsample would provide a crude check whether the AOI's behavior, and its rejection rate in particular, was in line with its normal performance (as reflected in the historical reject rate). The assumption that the inspections in the Tot-, Rej- and His-samples are comparable is an important one. However, in hindsight, we do not believe that the results from these 100 parts add much value in the evaluation of the AOI.

2.4.2 Statistical analysis

The parameters (α_j, δ_j) for the AOI and the operators are estimated by maximizing the likelihood (2.5) using the interior point algorithm. The fitted parameters are

	estimate (150-logistic)	*se*	estimate (150-loglgc)	*se*	estimate (149-loglgc)	*se*
δ_{AOI}	2.58	0.0098	2.55	0.0102	2.56	0.0102
IAP_{AOI}	0.0673	0.0095	0.0728	0.0101	0.0695	0.0100
IRP_{AOI}	0.0004	0.0001	0.0001	0.0000	0.0001	0.0000
δ_{opr}	3.37	0.0845	3.21	0.0617	3.22	0.0628
IAP_{opr}	0.2501	0.0254	0.0951	0.0379	0.0994	0.0386
IRP_{opr}	0.0004	0.0001	0.0000	0.0000	0.0001	0.0000

Table 2.2: Point estimates and standard errors based on the original analysis (150-logistic), loglogistic curves (150-loglgc) and after removing one anomalous part (149-loglgc). Standard errors calculated as in (2.6) and (2.7).

		Observed		Predicted (150-logistic)		Predicted (150-loglgc)	
		Rej-sample	Tot-sample	Rej-sample	Tot-sample	Rej-sample	Tot-sample
Number of rejections	0	0	99	2.90	99.4	0.12	99.45
	1	0	0	3.03	0.08	0.33	0.00
	2	1	0	3.39	0.04	0.72	0.01
	3	6	0	4.01	0.03	1.54	0.01
	4	6	0	5.05	0.03	3.53	0.02
	5	6	0	7.08	0.03	9.53	0.04
	6	21	0	12.7	0.05	31.8	0.12
	7	110	1	112	0.38	102	0.36

Table 2.3: Observed rejection frequencies (AOI only), and rejection frequencies predicted from a fitted logistic and loglogistic curve.

$\widehat{\alpha}_{\text{AOI}} = 26.69$ and $\widehat{\delta}_{\text{AOI}} = 2.582$ for the AOI and $\widehat{\alpha}_{\text{opr}} = 5.741$ and $\widehat{\delta}_{\text{opr}} = 3.369$ for the operators. Table 2.2 (leftmost columns) gives the corresponding values of $\widehat{IAP}, \widehat{IRP}, \widehat{\delta}$ and their standard errors.

To test goodness of the fit, we determine the number *df* of degrees of freedom of the G-statistic. For the Tot-sample (evaluated 7 times by the AOI) the possible response patterns for a part are $(0;7), (1;6), ..., (7;0)$, and a saturated model has $8 - 1 = 7$ *df*. For the His-sample we have $(0;1)$ and $(1;0)$ so 1 *df*. For the Rej-sample (7 repeats for the AOI, 3 for the operators) we have $8 \times 4 - 1 = 31$ *df*. In total, a saturated model has 39 *df*, while the fitted model has $2 \times J = 4$ *df*. Thus, the G-statistic has 35 *df*.

The G-test rejects the fit $\left(G = 127, p < 10^{-11}\right)$. As often with real data, there are a few issues, which are best explained by discussing the results of the AOI first. Table 2.3 (columns headed Observed) presents response pattern frequencies. The left column indicates how many of the 150 parts from the Rej-sample were rejected $0, 1, ..., 7$ times by the AOI. The next column shows the same for the 100 parts from the Tot-sample.

Although the results from the Tot-sample fit quite well, the fit is not that good

Rejections by AOI	Observed: Rejections by operators				Tot	Predicted (150-loglgc): Rejections by operators				Tot	Predicted (149-loglgc): Rejections by operators				Tot
0	0	0	0	0	99	0.12	0.00	0.00	0.00	99.5	0.19	0.00	0.00	0.00	99.5
1	0	0	0	0	0	0.33	0.00	0.00	0.00	0.00	0.46	0.00	0.00	0.00	0.01
2	1	0	0	0	0	0.72	0.00	0.00	0.00	0.01	0.91	0.00	0.00	0.00	0.01
3	6	0	0	0	0	1.54	0.00	0.00	0.00	0.01	1.74	0.00	0.00	0.00	0.01
4	5	0	0	**1***	0	3.52	0.00	0.00	**0.00***	0.02	3.58	0.00	0.00	0.00	0.02
5	5	1	0	0	0	9.47	0.01	0.01	0.03	0.04	8.68	0.00	0.00	0.01	0.04
6	18	0	1	2	0	30.2	0.27	0.34	1.05	0.12	27.0	0.19	0.22	0.60	0.10
7	93	2	3	12	1	81.6	2.59	3.56	14.6	0.36	85.0	2.84	3.65	14.0	0.37
	Historical rejection rate (AOI): 1,271 out of $I^{\text{His}} = 254,200$					Predicted rejection rate (AOI): 1,272 out of $I^{\text{His}} = 254,200$					Predicted rejection rate (AOI): 1,271 out of $I^{\text{His}} = 254,200$				

Table 2.4: Response pattern frequencies (observed and predicted from all data and after removing the results of one part)

for the Rej-sample. In particular, the model overestimates the number of parts with a small number of rejections, and underestimates the number of parts with 6 rejections. This can be seen from the raw differences between observed and predicted frequencies, or better, by calculating the Freeman-Tukey variance stabilized residuals: $-2.55, -2.62, -1.40, 0.97, 0.49, -0.32, 2.08, -0.15$ (for 0 to 7 rejects). Because the collation of observed to predicted frequencies suggests asymmetry in the characteristic curve, we fit loglogistic curves instead (as in (2.8)). The G-test confirms that this model fits better ($G = 42.4, df = 33, p = 0.13$). The estimated characteristics are in the middle columns of Table 2.2 (labeled 150-loglgc). The decision thresholds are obtained by solving $\widehat{q}_j\left(\widehat{\delta}_j\right) = 0.5$.

A second issue in the data is revealed by comparing the results for the AOI with those of the operators (Table 2.4). One part (marked with an asterisk) was rejected 3 (out of 3) times by the operators and only 4 (out of 7) times by the AOI. We consistently find that the AOI's decision threshold is stricter than that of the operators (Table 2.2), and consequently, there is no realization for misalignment Z for which this is a plausible response pattern (given the fitted model):

$$\max_z P_{\widehat{\theta}_{\text{AOI}},\widehat{\theta}_{\text{opr}}}\left[\boldsymbol{R}_i = \begin{pmatrix} 3 & 4 \\ 0 & 3 \end{pmatrix}\right] \approx 4.7 \times 10^{-5}.$$

We were unfortunately not in a position to inspect the part in question, and can only speculate that it may have had an abnormality or flaw that led the operators to reject it unanimously, but that was to some extent different from the usual misalignment problems, and therefore, not convincingly picked up by the AOI's algorithms. For its being an anomaly amid the results of the other 149 parts, we removed the results for this part from the analysis, and re-estimated the model.

The resulting parameter values of the loglogistic characteristic curves are $\widehat{\alpha}_{\text{AOI}} = 60.2$, $\widehat{\beta}_{\text{AOI}} = 1.26$ and $\widehat{\mu}_{\text{AOI}} = 2.54$ for the AOI and $\widehat{\alpha}_{\text{opr}} = 7.32$, $\widehat{\beta}_{\text{opr}} = 3.75$ and

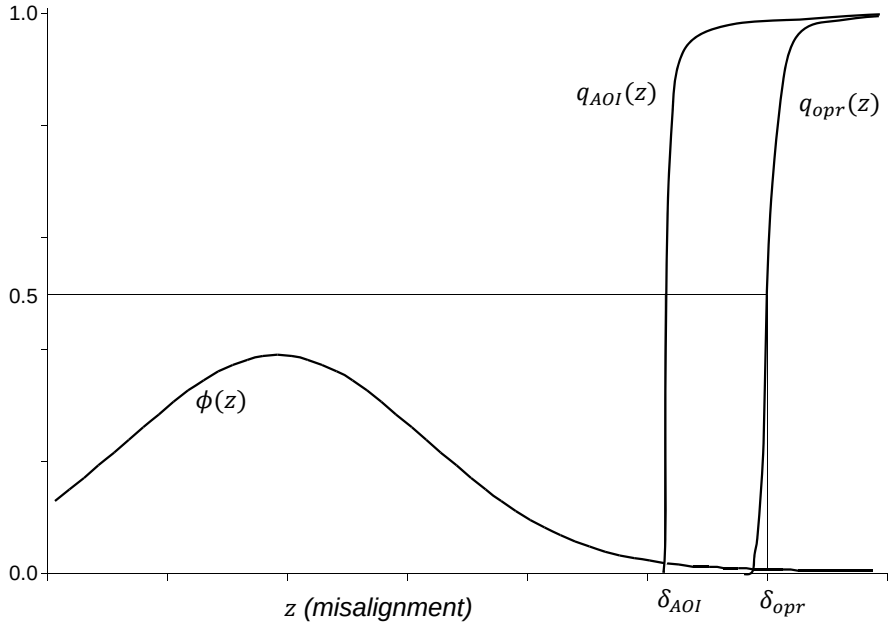


Figure 2.4: Fitted loglogistic characteristic curves of the AOI and operators.

$\widehat{\mu}_{\text{opr}} = 3.09$ for the operators. The fit has improved substantially, as is confirmed by the G-test ($G = 28.3, df = 33, p = 0.70$). Figure 2.4 shows the fitted characteristic curves of the AOI and the operators.

Table 2.2 (rightmost columns, labeled 149-loglgc) presents the final estimated characteristics of the inspections. The key results of the analysis are as follows.

- The AOI's repeatability can be characterized as: $\widehat{IAP}_{\text{AOI}} = 0.0695$ and $\widehat{IRP}_{\text{AOI}} = 0.0001$.
- The reproducibility of the operators can be characterized as $\widehat{IAP}_{\text{opr}} = 0.0994$ and $\widehat{IRP}_{\text{opr}} = 0.0001$.
- The systematic difference between the AOI and the operators is fairly substantial: $\widehat{\delta}_{\text{AOI}} - \widehat{\delta}_{\text{opr}} = -0.66$ (the AOI being stricter than the operators).

Further practical ramifications for the AOI are as follows. We could interpret the company's stance that the operators are the de facto standard by setting $USL = \widehat{\delta}_{\text{opr}}$.

This allows the calculation of the AOI's misclassification probabilities: $\widehat{FAP}_{\text{AOI}} = P\left[Y = 1 | Z > \widehat{\delta}_{\text{opr}}\right] = 0.0064$, and $\widehat{FRP}_{\text{AOI}} = P\left[Y = 0 | Z \leq \widehat{\delta}_{\text{opr}}\right] = 0.0044$. The estimated fraction of nonconforming parts in the stream of parts accepted by the AOI is $P\left[Z > \widehat{\delta}_{\text{opr}} | Y = 1\right] = 3.79 \times 10^{-6}$, and the estimated fraction of conforming parts in the stream of rejects is $\widehat{P}\left[Z \leq \delta_{\text{opr}} | Y = 0\right] = 0.8822$ (these fractions are 45.9×10^{-6} and 0.0999 for the operators' inspections).

2.4.3 Discussion and appraisal of the car parts MSA study

For the results reported above, we must make an important reservation, that they only hold if the complication that manifested itself in the one removed part is under control. It is an unsatisfactory situation that we cannot inspect the part in question. Instead, we cannot do more than recommend keeping the AOI under surveillance for a couple of weeks more. Any part found during this period that is accepted by the AOI, but rejected by the operators, should be closely inspected, as this inspection result is singular and indicative of a specific failure mode that the AOI's algorithms do not detect reliably. Also, those parts in the MSA experiment for which the AOI's 7 appraisals were not in agreement could be inspected for clues to improve the AOI's inspection algorithms.

Once this complication is judged under control, we can conclude that the AOI has no worse reproducibility than the team of operators, but it turns out to act on a stricter decision threshold. Following the company's stance that the operators' decisions constitute the de facto standard, this implies that the decision threshold of the AOI should be adjusted. This could be accomplished by selecting from the Rej-sample the 5 parts that have been rejected 1 or 2 times (out of 3) by the operators, and 7 times by the AOI. Using these parts as a training set, the AOI should be adjusted until it rejects these parts about half the time. After effective adjustment, the AOI is expected to perform equally well as the operators.

Even though the combination of samples that we obtained is not optimal, and despite the issues encountered during the analysis, we believe that the analysis is reasonable. The standard errors in Table 2.2 reveal that the estimates are reasonably precise (except for the estimated IAP_{opr}, which we discuss below). The results of various alternative analyses (150-logistic, 150-loglgc, 149-loglgc) are very close (again with the exception of $\widehat{IAP}_{\text{opr}}$). The $\widehat{IAP}_{\text{opr}}$ should be taken with fairly large confidence margins ($se = 0.0386$, which is large in view of the estimated value of $\widehat{IAP}_{\text{opr}} = 0.0994$). The standard error would have been smaller if the Rej-sample had been sampled from the population of items rejected by the operators instead of the AOI. This is because the AOI turns out to be substantially more strict than the operators (Figure 2.4), and consequently, most of the items rejected by the AOI have Z-values close to δ_{AOI}

but to the left of δ_{opr}. This in turn implies that there is limited information in the sample that we obtained for fitting the middle and right part of q_{opr}. In hindsight, sampling 100 parts from the stream of items rejected by the operators would have been better than the 100 parts sampled from the total population (the Tot-sample).

Besides the three analyses summarized in Table 2.2, we also fitted a model with a loglogistic (asymmetric) curve for the AOI, and a logistic (symmetric) curve for the operators. This alternative analysis and our final fit (the 149-loglgc analysis) are equivalent in terms of goodness of fit ($G = 28.6$ instead of $G = 28.3$), and also the estimates are very close. The largest difference is for $\widehat{IAP}_{\text{opr}}$, which is estimated as 0.0774 instead of 0.0994. Note that this difference is small compared to the standard error of this estimate.

2.5 Discussion

In the recent academic literature, it is generally assumed that the traditional notion of constant misclassification probabilities across items and situations is too simplistic for most applications (see the references in the introduction). Current literature explores various ways to accommodate this variability in the statistical models underlying the evaluation approaches. In our modeling, we have attributed the variability of the misclassification probabilities to the degree to which the condition is present that the inspections aim to detect. Thus, variability in the misclassification probabilities is linked by a characteristic curve to the stochastic properties of a random variable Z, misalignment in the case study. This type of modeling is explored further in De Mast et al. (2014).

Two other approaches in the literature are closely related. The traditional approach for MSA under absence of a gold standard is to fit a latent class model (references given in the introduction), which assumes that the rejection probabilities are constant in the subpopulations of good and defective items. However, this is now generally considered an implausible assumption in almost every situation (e.g. De Mast et al., 2011).

Danila et al. (2012) propose an alternative approach, based on a random effects model. That model, like the model presented in this paper, also allows nonconstant error rates. However, variability in the error rates is not attributed to specific factors by means of link functions, but instead, an item's misclassification probabilities are assumed to be realizations of two probability distributions on the unit interval, one for good and one for defective items. An advantage of that model is that it allows the measurement outcomes to be affected by more than one property of the items. However, it does not distinguish between variability induced by properties of the items (such as misalignment) and variability induced by properties of the inspection system (such as the characteristic curve). We think that an advantage of the approach based

on latent trait models, is that it gives more informative results and can also be naturally extended to multiple appraisers or multiple inspection systems. Note how the fitted characteristic curves of the AOI and the operators improve the understanding of the reliability of the inspections and inspire ways for improving the AOI's performance. An objective of later chapters is to produce more detailed recommendations where one or the other approach is more promising. See Chapters 3 and 4.

A difficult problem in designing MSA studies, noted in the literature on the evaluation of industrial tests and medical tests as well, is that a sample from the total population is often substantially unbalanced. In terms of the characteristic curve model in this chapter, the part of the characteristic curve q that is close to, or to the right of, δ is typically in the remote right tail of the density ϕ of Z. As a consequence, simply sampling from the total items population provides an ineffective basis for estimating ϕ and q simultaneously. We believe that our approach, which combines samples from multiple sources, and takes the origin of each sample into account in the estimation procedure, offers an effective solution to this problem, as we have demonstrated for the specific case that we have described. Simulations show that the estimators $\widehat{IAP}$ and $\widehat{IRP}$ have the highest precision if only rejected items are included in the MSA experiment and the data are supplemented with a historical dataset. Furthermore, simulations show that this procedure is robust to certain forms of misspecification: even if the distribution of Z has heavy tails or is asymmetric, or if the Z is multi-dimensional, the estimators perform reasonably well in terms of bias and precision.

Chapter 3

Estimation of FAP and FRP without gold standard: mathematical analysis

The first goal of this thesis is to obtain a better understanding of models for binary measurement, and to find the implications of an unavailable gold standard. The previous chapter did so using a single case study. This chapter offers a general modeling framework in terms of rejection probabilities and rejection counts. Models for binary measurement systems in literature can be written in terms of this framework. We use this model to arrive at the general conclusion that estimation of the false acceptance and rejection probabilities (FAP and FRP) is problematic in the absence of a gold standard.

The second goal of this thesis is to develop a reliable method for estimating measurement error. This chapter takes a first step by introducing the random components of the FAP and FRP, and shows that these are the only components of measurement error that can generally be estimated without gold standard. A well-grounded method for estimation of these random components will be presented in Chapter 5.

3.1 Introduction

In a measurement system analysis (MSA) study, one evaluates how reliably a measurement result Y reflects the true state X. For binary measurements, this reliability is expressed using misclassification probabilities (AIAG, 2003; De Mast et al., 2014). Let $X = 1$ denote that a product conforms to specifications, and $X = 0$ that it does

not. $Y = 1$ denotes the acceptance of a product, and $Y = 0$ denotes a rejection. The misclassification probabilities are:

$$FAP = P[Y = 1|X = 0] \text{ (False Acceptance Probability)}$$
$$FRP = P[Y = 0|X = 1] \text{ (False Rejection Probability)}$$

A situation in which it is possible to obtain X is called a "gold standard available" situation. A gold standard is an authoritative measurement procedure that is accepted by convention to give a faithful representation of the true state. In such case, estimation of *FAP* and *FRP* is straightforward. Without a gold standard, however, more elaborate modeling becomes necessary. The traditional way is to fit a latent class model, that treats the unobservable true value as a latent dichotomous variable (Hui and Walter 1980; Boyles 2001; Van Wieringen and De Mast 2008; Danila et al. 2010; Beavers et al. 2011 and many others). Such models imply that the *FAP* (*FRP*) is the same for all defective (conforming) items. However, in recent years this assumption has been discredited as generally unrealistic, since typically, some defective items have a larger *FAP* than others (and analogously for conforming items and their *FRP*), and also the conditions in which tests are performed induce variability in the *FAP* and *FRP*. It has been shown that estimates for *FAP* and *FRP* from the simplistic traditional latent class model are often substantially biased (Vacek, 1985; Torrance-Rynard and Walter, 1997; Irwig et al., 2002; De Mast et al., 2011).

Various more complex latent variable approaches have been suggested to accommodate variable acceptance probabilities of truly defective items, and variable rejection probabilities of truly conforming items. These approaches are based on rather different premises and model assumptions, and their robustness against model misspecification has been seriously questioned (Pepe, 2003; Albert and Dodd, 2004; Akkerhuis et al., 2016b). To make matters worse, the true stochastic mechanisms are generally unknown in real applications, and from the binary observations violations of model assumptions are typically difficult to diagnose (Albert and Dodd, 2004). Influential texts in medical testing (Pepe, 2003; Albert and Dodd, 2004) go so far as to warn against the use of latent variable modeling, and are skeptical whether the reliability of binary tests can be established in the absence of a gold standard.

The ubiquity of binary tests in industry, medicine and elsewhere, and the often serious ramifications of classification errors of such tests, are a strong motivation to develop reliable methods for binary MSA. For the common situation that a gold standard is unavailable, convincing progress has been lacking. So far, there is no approach for determining the *FAP* and *FRP* whose effectiveness and reliability are widely accepted and supported by strong evidence.

This chapter aims to take a step back and reconsider the problem. Rather than proposing yet a new method, we put forward a novel statistical modelling framework. This modelling framework supports a number of remarkable conclusions, that clarify

the stubborn problems in developing effective methods. We demonstrate that generally, *FAP* and *FRP* are unidentifiable quantities when no gold standard is available, and therefore, that their estimation is inherently problematic. This mere finding goes a long way in explaining the encountered difficulties in developing reliable methods. We also show that the *FAP* and *FRP* can be decomposed into a systematic and a random component. And finally, we show that these random error components can in general be estimated. Granting that knowledge of the *FAP* and *FRP* would be preferable, but accepting that this may be an ambition that is not generally attainable without gold standard, we propose to shift focus from estimation of *FAP* and *FRP* to estimation of the random errors, and we motivate that these are quite useful results in themselves for many applications. Throughout this chapter, we illustrate our proposals with a real-life application from the inspections of car parts.

3.2 Statistical modeling

3.2.1 Rejection counts U_i and rejection probabilities R_i

In a regular application of a test in quality inspection, items $i = 1, 2, \ldots$ are tested once, which gives the test results $Y_i = 1$ (accepted) or $Y_i = 0$ (rejected). For determining the reliability of the test, one performs an MSA study, in which a sample $i = 1, \ldots, I$ of items are tested multiple times, giving the repetitions $k = 1, \ldots, K$. The raw data of the MSA study are denoted $Y_{ik} \in \{0, 1\}$, and they can be summarized in the per item rejection counts $U_i = \sum_{k=1}^{K} (1 - Y_{ik}) \in \{0, \ldots, K\}$. The main idea of our modelling approach is that without gold standard, this rejection count is the only observable property of the items, and therefore, the model's central elements are the observed rejection fractions U_i/K and their expectations: the items' rejection probabilities R_i.

An items' true state $X_i \in \{0, 1\}$ is assumed unobservable, and the fraction

$$p = P[X_i = 0] \tag{3.1}$$

of defectives in the population of items is thus unknown. The probability that item i is rejected is $R_i = P[Y_{ik} = 0]$. In the traditional latent class model, $R_i = 1 - FAP$ for all defective items and $R_i = FRP$ for all conforming items (with *FAP* and *FRP* two constants). Recent consensus is that this rendering is too simplistic. For example, the dichotomous true state X_i almost always masks an underlying continuum where some defective items are more severely nonconforming than others, and it is not realistic to assume that all of them have identical *FAP* (and analogously for conforming items and their *FRP*). In addition, conditions under which the tests are performed may have an effect on the test results, inducing further variability in the rejection probabilities.

We allow variability by modelling an item's rejection probability R_i as a random effect. In the population of items, the distribution of rejection probabilities is $F_R(r) = P[R_i \leq r], r \in [0,1]$. We assume that conditional on an item's rejection probability, repeated test results are independent. Thus, conditional on the event $\{R_i = r_i\}$ the $Y_{i1} \ldots, Y_{iK}$ are i.i.d. Bernoulli $(1 - r_i)$ variables, and the rejection counts U_i have a binomial (K, r_i) distribution.

The distribution of rejection probabilities $F_R(r) = p \times F_R^0(r) + (1-p) \times F_R^1(r)$ is a mixture of two distributions, with $F_R^0(r) = P[R_i \leq r | X_i = 0]$ and $F_R^1(r) = P[R_i \leq r | X_i = 1]$. Acknowledging that the probability of a misclassification varies per item, we define the *FAP* and *FRP* in terms of the means of these component distributions:

$$FAP = E[1 - R_i | X_i = 0];$$
$$FRP = E[R_i | X_i = 1].$$

3.2.2 Motivation by a mechanistic model

To motivate and demonstrate the generality of the presented model, we show how plausible mechanisms result in the given model. One common origin of the variability in misclassification probabilities is that the true state X_i is related to a continuous property $Z_{0,i}$. Chapter 2 presents a real example concerning the assembly of car parts, which are accepted $\{Y = 1\}$ or rejected $\{Y = 0\}$ by an optical quality inspection system. The parts' true states are determined by the misalignment Z_0 of a clip and the pad to which it is to be fastened: a part is considered conforming $(X = 1)$ if misalignment is below an upper specification limit *USL* $(Z_0 \leq USL)$, and defective otherwise.

For establishing the reliability of the inspection system, a gold standard was not available. Because of the very small scale and uneven surface of the parts, misalignment could not be measured directly and there is no stated value for *USL*. The optical inspection system uses a proprietary algorithm and has been calibrated to some extent on the basis of a training set. This training set consisted of parts that were considered conforming or defective by consensus among engineers. Given the far from negligible disagreement among engineers regarding some of the parts, these judgments could not be used as a gold standard however.

When the true state X is a dichotomization of a continuous property Z_0, rejection probabilities typically vary from $R_i = 0$ (parts with no misalignment) up to $R_i = 1$ (substantial misalignment): Figure 3.1 illustrates the idea. The relation between an item's R_i and $Z_{0,i}$ could be modeled as a characteristic curve, such as a logistic curve:

$$q(z_0) = P[Y_{ik} = 0 | Z_{0,i} = z_0] = (1 + \exp\{-\alpha(z_0 - \delta)\})^{-1} \tag{3.2}$$

Parts with misalignment $Z_{0,i} = \delta$ have rejection probability $R_i = q(\delta) = 0.5$. For parts with $Z_{0,i} \leq \delta$, the most likely test result is $Y_i = 1$, as is $Y_i = 0$ for parts with $Z_{0,i} > \delta$. The distribution of rejection probabilities is

$$F_R(r) = P[q(Z_{0,i}) \leq r] = F_{Z_0}\left(q^{-1}(r)\right),$$

with $F_{Z_0}(z)$ the distribution of misalignment in the population of items. Figure 3.2, which could be imagined on the y-axis of Figure 3.1, shows the density f_R of rejection probabilities implied by the logistic curve q and normal density f_{Z_0} in Figure 3.1. The mean probabilities of false acceptance and rejection are

$$FAP = \int_{USL}^{\infty} (1 - q(z)) f_{Z_0}(z) \, \mathrm{d}z / p,$$
$$FRP = \int_{-\infty}^{USL} q(z) f_{Z_0}(z) \, \mathrm{d}z / (1 - p),$$

with $p = P[Z_{0,i} > USL]$ as per (3.1). Note that, clearly, the assumption of the traditional latent class model is violated, since it is not true that all defective parts have $R_i = 1 - FAP$ and all conforming parts have $R_i = FRP$ (with FAP and FRP two constants).

Besides the fact that a continuum underlies the measurand, a second common cause of the variability in misclassification probabilities is the effect of nuisance variables $Z_1, \ldots, Z_M$. These could be properties of the items or the test conditions and appraisers, such as light conditions and appraiser fatigue. Their effects on the test results can be incorporated in the characteristic curve:

$$q(z_0; z_1, \ldots, z_M) = P[Y = 0 | Z_0 = z_0, Z_1 = z_1, \ldots, Z_M = z_M]$$

(this is similar to what is called the "measurement model" in metrology; see JCGM, 2008). An example of such a characteristic function is given in De Mast et al. (2014). We will elaborate on the effects of nuisance variables and their ramifications for the design of the MSA study in a later section.

3.2.3 Estimation of FAP and FRP is problematic

Avoiding the unwarranted simplicity of the traditional latent class approach, where rejection probabilities are considered constant in the subpopulations of conforming and defective items, Danila et al. (2012) give an approach that allows variable rejection probabilities. The distribution $F_R^0(r)$ of rejection probabilities in the population of

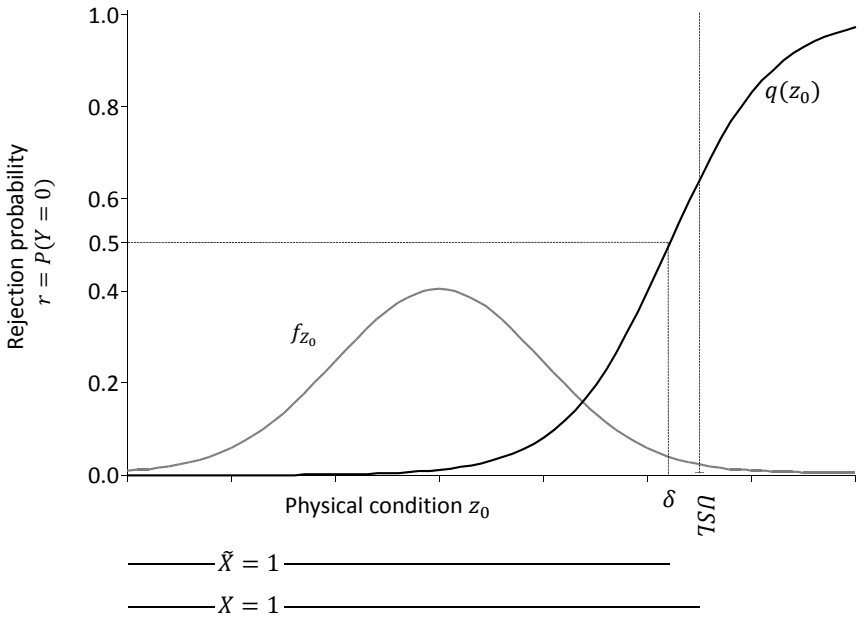


Figure 3.1: Density f_{Z_0} of misalignment Z_0, and a characteristic curve $q(z_0) = P[Y = 0|Z_0 = z_0]$. The upper specification limit *USL* demarcates the range of z_0 values for which items are considered conforming ($X = 1$).

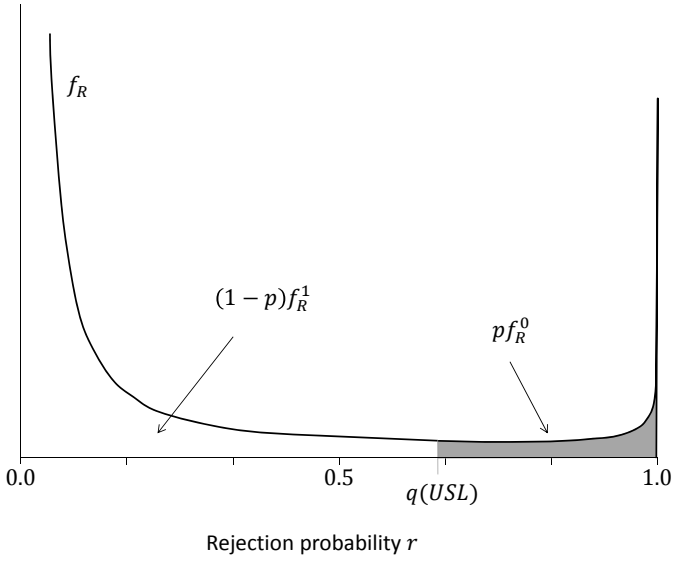


Figure 3.2: Density f_R implied by f_{Z_0} and q in Figure 3.1, and split into $p \times f_R^0$ (gray) and $(1-p) \times f_R^1$ (white).

defective items is a beta distribution with mean $1-\mu_0 = 1-FAP$ and dispersion ϕ_0, and the distribution $F_R^1(r)$ in the population of conforming items is a beta distribution with mean $\mu_1 = FRP$ and dispersion ϕ_1. The distribution $F_R(r) = P[R \leq r]$ in the entire items population is a mixture of the two, with density:

$$f_R(r) = (1-p) \times \frac{r^{\mu_1 \frac{1-\phi_1}{\phi_1}} (1-r)^{(1-\mu_1)\frac{1-\phi_1}{\phi_1}-1}}{\text{Beta}\left(\mu_1 \frac{1-\phi_1}{\phi_1}, (1-\mu_1)\frac{1-\phi_1}{\phi_1}\right)} + p \times \frac{r^{(1-\mu_0)\frac{1-\phi_0}{\phi_0}-1} (1-r)^{\mu_0 \frac{1-\phi_0}{\phi_0}-1}}{\text{Beta}\left((1-\mu_0)\frac{1-\phi_0}{\phi_0}, \mu_0 \frac{1-\phi_0}{\phi_0}\right)}. \quad (3.3)$$

The parameters are estimated by maximum likelihood, which gives estimates for *FAP* and *FRP*. Variations on this idea are the methods proposed by Qu et al. (1996) and Albert et al. (2001).

Simulation studies (e.g. Albert and Dodd, 2004; Akkerhuis et al., 2016b and Chapter 4) show that such methods only work reliably under quite restrictive circumstances, when the variability in rejection probabilities is relatively small. Further, it turns out that such methods easily derail under realistic violations of model assumptions, resulting in biases in the estimated *FAP* and *FRP* up to and exceeding their true values.

Our model, introduced above, demonstrates in an elegant and simple manner that it is not plausible that this poor performance can in general be improved, and that the estimation of *FAP* and *FRP* is inherently problematic when no gold standard is available. Remember that the true state X_i is unobservable, and therefore, that it is not possible to infer which of the observed rejection counts U_i belong to defective parts ($X_i = 0$) or conforming parts ($X_i = 1$). We show that it is possible to infer f_R, but that without additional information or (arbitrary) restrictions its decomposition into f_R^0 and f_R^1 is not identifiable. Namely, whether f_R is the mixture of two beta distributions mentioned above or another function, the likelihood of the observed rejection count of item i is

$$P[U_i = k] = \binom{K}{k} \int_0^1 r^k (1-r)^{K-k} f_R(r)\, \mathrm{d}r$$

(and analogously for the total likelihood $P[U_1 = k_1, \ldots, U_K = k_K]$). This likelihood depends on f_R, but not on the components f_R^0 and f_R^1 and $p = P[X_i = 0]$. Put differently, each combination of f_R^0, f_R^1 and p for which $p \times f_R^0 + (1-p) \times f_R^1 = f_R$, produces the same likelihood $P[U_i = k]$. For this reason, attempting to fit f_R^0 and f_R^1 by maximizing the likelihood will in general produce results that are driven by

(arbitrary) assumptions on the exact forms of f_R^0 and f_R^1, but that fail to reflect the decomposition of f_R on the basis of the $X_i = 0$ and $X_i = 1$ conditions. To illustrate: increasing or decreasing the value of *USL* in Figure 3.1 shifts the demarcation between $p \times f_R^0$ and $(1-p) \times f_R^1$ in Figure 3.2 to the left or right, but f_R, the mixture of the two, does not change and consequently, neither does the likelihood. Thus, it is not possible to infer the location of the demarcation from the data by maximum likelihood. In effect, this demonstrates that *FAP* and *FRP*, being the means of the unidentifiable components f_R^0 and f_R^1, are unidentifiable parameters. This mathematical result, combined with the results of the simulation studies in Chapter 4, motivates our conclusion that the development of reliable methods for estimating *FAP* and *FRP* while a gold standard is unavailable, is an unattainable ambition.

3.3 Estimation of random error

In MSA studies for numerical measurements, measurement error is decomposed into systematic error (sometimes called bias, trueness or accuracy) and random error (sometimes called precision). This distinction is recognized in international standards in metrology such as ISO (1995), AIAG (2003) and JCGM (2008). Systematic error is the discrepancy between the average measurement outcome for an item and the item's true value. Systematic error is dealt with by calibrating the measurement system. It can only be established by using a gold standard or reference standard, which is analogous to our conclusion in the previous section for binary measurements. The random measurement error quantifies the variability of measurements around the mean result for an item. It reflects how consistent measurement results are, but ignores whether results are systematically off. Whether a gold standard is available or not, it can be determined using a gauge repeatability and reproducibility study or similar approach (Vardeman and Van Valkenburg, 1999; AIAG, 2003).

We demonstrate now that also for binary tests, the total errors (*FAP* and *FRP*) can be decomposed into systematic error and random error, and further, that the latter can in general be estimated without gold standard. Let X_i, as before, be the true state of an item i, and let $\tilde{X}_i$ denote the modal (most likely) test result. Note that $\tilde{X}_i = 0$ if $R_i > 0.5$, and $\tilde{X}_i = 1$ if $R_i \leq 0.5$. We have

$$FRP = P\left[Y = 0 | X = 1\right] = P\left[Y = 0, \tilde{X} = 0 | X = 1\right] + P\left[Y = 0, \tilde{X} = 1 | X = 1\right].$$

The first term on the right-hand side is the fraction of the *FRP* where results are systematically off (that is, $\tilde{X} \neq X$). The second term is the fraction that is due to random deviations of results from the modal result ($Y \neq \tilde{X}$). The random component

decomposes further into

$$P\left[Y=0,\tilde{X}=1|X=1\right]=P\left[Y=0|\tilde{X}=1\right]\times\frac{P\left[\tilde{X}=1\right]}{P\left[X=1\right]}-$$
$$P\left[Y=0|\tilde{X}=1,X=0\right]\times\frac{P\left[\tilde{X}=1,X=0\right]}{P\left[X=1\right]}.$$

The first term we call the Inconsistent Rejection Probability (*IRP*):

$$P\left[Y_{ik}=0|\tilde{X}_i=1\right]=E\left[R_i|R_i\leq 0.5\right]=IRP.$$

IRP is the probability that a result randomly deviates from the most likely result for the item, and it is a measure for the random variability of test results. As a component of the total error *FRP*, it is corrected by the probability $P\left[Y=0|\tilde{X}=1,X=0\right]$ that a random deviation cancels out a systematic deviation, in which case there is no error (i.e., $Y=X$). Note that this probability is zero for numerical measurements as a consequence of continuity, but it is generally nonzero for nonnumerical measurements.

In summary: the total error *FRP* is decomposed into fractions where results are systematically or randomly off, and the latter equals the *IRP* less the probability that a random deviation cancels out a systematic deviation. Similarly, the *FAP* may be decomposed into a systematic part and a random component, the latter being the Inconsistent Acceptance Probability (*IAP*):

$$P\left[Y=1|\tilde{X}=0\right]=E\left[1-R|R>0.5\right]=IAP.$$

Applying these concepts in the car parts example, we find that

$$\begin{aligned} IAP &= P\left[Y=1|Z_0>\delta\right], \\ IRP &= P\left[Y=0|Z_0\leq\delta\right]. \end{aligned} \tag{3.4}$$

Here, δ is the misalignment value for which $q\left(\delta\right)=0.5$. In case the characteristic curve is a logistic curve (as in (3.2)), then δ is the curve's inflection point. For items with misalignment $Z_0\leq\delta$, the value $Y=1$ is the most likely test result, and therefore, δ can be seen as the rejection threshold that the system applies. If an item with $Z_0\leq\delta$ is rejected, this is inconsistent with the system's rejection threshold (and vice versa if an item with $Z_0>\delta$ is accepted). Defined as in (3.4), the inconsistent rejection and acceptance probabilities were introduced earlier by De Mast et al. (2011) and Erdmann et al. (2016a) and borrowed from Item Response Theory (Lord, 1980).

It is obvious that the assessment of a test in terms of both the systematic and the random error is desirable. But for lack of a gold standard, the determination of

the random errors *IAP* and *IRP* is generally all one can hope for. We do believe that knowing a test's random errors is generally useful in itself for evaluating a test. First, the *IAP* and *IRP* show the extent to which the test yields reproducible results. If *IAP* and *IRP* are poor, this means that the test is very uninformative and gives highly random results (noise). Whether the test can be calibrated or not, this invalidates the test. If *IAP* and *IRP* are good, then at least the test is capable of discriminating between worse and better items, which may be sufficient for many applications. And if *IAP* and *IRP* turn out to be really good and $f_R(r)$ is almost zero for values around $r = 0.5$, then even a rough calibration assures that *IAP* and *IRP* approximate *FAP* and *FRP*.

3.3.1 Estimation of *IAP* and *IRP*

Estimation of *IAP* and *IRP* is relatively straightforward. First, one fits a density $\hat{f}_R$ to the observed rejection counts U_i. One option is the mixture of beta distributions in (3.3) proposed by Danila et al. (2012). The five parameters (p, μ_0, μ_1, ϕ_0 and ϕ_1) are estimated by maximum likelihood. In its original application, the μ_0 and μ_1 of the fitted density are used as estimates for *FAP* and *FRP*. Ignoring this original purpose, we can determine *IAP* and *IRP* from the fitted density by

$$\begin{aligned} IAP &= E\left[1 - R_i | R_i > 0.5\right] = \frac{\int_{0.5}^{1} (1 - r)\, \hat{f}_R(r)\, \mathrm{d}r}{\int_{0.5}^{1} \hat{f}_R(r)\, \mathrm{d}r}, \\ IRP &= E\left[R_i | R_i \leq 0.5\right] = \frac{\int_{0}^{0.5} r \hat{f}_R(r)\, \mathrm{d}r}{\int_{0}^{0.5} \hat{f}_R(r)\, \mathrm{d}r}. \end{aligned} \tag{3.5}$$

This procedure's effectiveness was demonstrated in a wide range of scenarios (Akkerhuis et al., 2016b). Note that the effectiveness of this procedure does not hinge on the fitted f_R being a mixture of two densities f_R^0 and f_R^1. One could consider fitting any class of distributions on $[0, 1]$ flexible enough to approach the distribution of rejection probabilities under a relevant range of scenarios. Akkerhuis et al. (2016a) present a more parsimonious and yet more flexible alternative, also offering recommendations for sample sizes (see Chapter 5 of this thesis).

3.3.2 Estimation of *IAP* and *IRP* in the presence of nuisance variables

An important source of variability in the rejection probabilities is the effect of nuisance variables such as variable light conditions, appraiser fatigue, time pressure, motivation and other variation in the conditions under which tests are done. To explain the idea, we limit the discussion to a single nuisance variable Z_1. It may vary not only over items, but also over repeated judgments of the same item, and therefore, we denote by $Z_{1,ik}$ the value of the nuisance variable during the kth judgment of item i.

As before, we denote by Z_0 the continuous property that determines the true state, such as misalignment in the car parts example. This gives the characteristic curve

$$q\left(z_0, z_1\right) = P\left[Y_{ik} = 0 | Z_{0,i} = z_0, Z_{1,ik} = z_1\right]$$

We assume that, conditional on $Z_{0,i}$ and $Z_{1,ik}$, repeated judgments are independent.

The dashed curves in Figure 3.3 are $q\left(z_0, Z_1\right)$ for three realizations of the nuisance variable Z_1; as one will observe, the probability of rejection not only depends on the relevant property Z_0, but also on the nuisance variable. An extreme case of what could happen if the experimenter does not design her MSA experiment carefully, is that all $I \times K$ tests are done under a single realization. If this realization is, say, $Z_1 = z_1^1$, the number of rejections is relatively high, while the opposite is true when $Z_1 = z_1^3$. Thus, the test's consistency under realistic variation in conditions may be much worse than implied by the estimated *IAP* and *IRP*.

The ideal way to deal with nuisance variables is of course to eliminate their effect in the test protocol, for example, by introducing countermeasures that keep nuisances constant. If this is not possible, we propose to handle nuisance variables by experimental randomization. This means that each of the K repeats per item is done under a new realization of the nuisance variables, and such that these realizations are representative for normal circumstances. This way, the effects of nuisance variables are treated as a source of random error, and ensures that their impact on the reproducibility of test results is reflected in the *IAP* and *IRP*. Let

$$\bar{q}\left(z_0\right) = \int q\left(z_0; z_1\right) f_{Z_1}\left(z_1\right) \mathrm{d}z_1$$

represent the mean response of the test to the relevant variable Z_0 under varying conditions Z_1 (the solid curve in Figure 3.3). We have $\bar{R}_i = \bar{q}\left(Z_{0,i}\right)$ is the mean rejection probability of item i under varying nuisance conditions, and

$$\begin{aligned} IAP &= 1 - E\left[\bar{R}_i | \bar{R}_i > 0.5\right], \\ IRP &= E\left[\bar{R}_i | \bar{R}_i \leq 0.5\right]. \end{aligned}$$

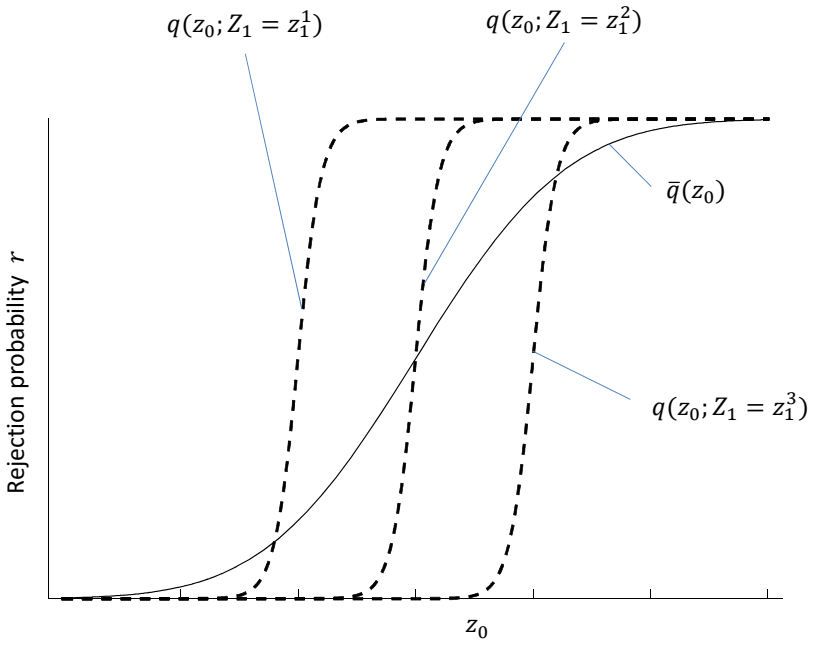


Figure 3.3: Characteristic curves $q\,(z_0, Z_1)$ for three realizations of Z_1 (dashed lines), and average characteristic curve $\bar{q}\,(z_0) = \int q\,(z_0, z_1)\, f_{Z_1}\,(z_1)\,\mathrm{d}z_1$.

When experimental randomization is applied correctly, the observed rejection fractions estimate $E\left[R_i | Z_{i,0} = z_{i,0}\right] = \bar{R}_i$, and the fitted distribution function approximates $f_{\bar{R}}(r)$. Taking the expected values under the conditions $\bar{R}_i > 0.5$ and $\bar{R}_i \leq 0.5$ gives estimates for $1 - IAP$ and IRP as defined above.

3.3.3 Design of the MSA study

Presenting and motivating detailed prescriptions for MSA studies and a method for analyzing the results require more space than we have here. We briefly mention some results published in Erdmann (2012), Akkerhuis (2012) and other recent papers. In Chapter 5, we will provide detailed prescriptions for the robust estimation method that we propose.

Erdmann et al. (2016a) and Akkerhuis et al. (2016a) recommend to take conditional samples for estimation of *IAP* and *IRP*; Danila et al. (2008, 2010, 2012, 2013) have done so for estimation of *FAP* and *FRP*. This means that the sample of I items is not taken from the population of items, but instead, is taken randomly from the population of rejected items. Given the typically very low rejection rates in industrial processes, a sample from the population of items would contain virtually no parts with $R_i > 0.5$, and consequently, estimation of *IAP* in particular would be imprecise (the abovementioned papers substantiate this intuitive claim with evidence from simulation studies). In the stream of rejected items, to the contrary, values of R_i are distributed more evenly. To obtain estimates for the total population of items from such sample, one uses the known relation between the distribution F_R in the items population, and $F_R^{\text{Rej}}(r) = \int_0^r f_R(t)\, t\mathrm{d}t \Big/ \left(\int_0^1 f_R(t)\, t\mathrm{d}t \right)$, which is the distribution in the subpopulation of rejected items. The abovementioned papers also show how the precision of estimates can be further improved by incorporating a historical estimate of the rejection rate in the estimation algorithm. With this set-up, a sample size of $I = 100$ to 150 items and $K = 7$ repetitions results in reasonable standard errors (Akkerhuis et al., 2016a).

In the car parts example, 150 parts were randomly taken from the stream of rejected parts, and these were judged 7 times by the optical inspection system. In addition, it was given that the reject fraction was 0.0050 in 254, 200 recently inspected parts. Fitting the model in (3.3) by maximum likelihood, and using (3.5), yields

$$\widehat{IAP} = 0.0778, \text{se}_{\widehat{IAP}} = 0.0107,$$
$$\widehat{IRP} = 0.0001, \text{se}_{\widehat{IRP}} = 0.0001$$

(data and details can be found in Akkerhuis et al., 2016a and are discussed in Chapter 5). The estimated *IAP* and *IRP* indicate that the optical inspections are reasonably

repeatable for parts with a large rejection probability, and very repeatable for parts with a low rejection probability. In practical terms, this means that the inspection system's discriminative capability is sufficient for discerning better from worse parts. Note how this result is quite useful in itself. Calibrating the system without gold standard remains problematic, and a rough calibration by a training set based on consensus among engineers may be the best one can do.

Originally, the model in (3.3) was intended for the estimation of the FAP and FRP (Danila et al., 2012). Applying this approach yields $\widehat{FAP} = 0.1108$ and $\widehat{FRP} = 0.0000$. Theoretically, 0.1108 is the estimated fraction of defective parts that are falsely accepted (and analogously for the fraction of conforming parts falsely rejected). But with conformance being an unobservable and only vaguely defined concept for this process without numerical specifications, we find it difficult to see how the concept of a false acceptance, and hence the produced estimate for FAP, can be related back to a relevant property of the inspection system. We are inclined to believe that instead, such values are artifacts from the model assumptions.

3.4 Conclusions

The determination of the error probabilities of binary tests when a gold standard is not available has been a stubborn problem, and currently, no reliable methods are available to the practitioner. This chapter introduces a statistical model in terms of rejection probabilities R_i and observed rejection counts U_i. We motivate the model by demonstrating how it naturally arises for plausible mechanisms that induce variability in the rejection probabilities. In elegant and simple manner, the model allows us to clarify the problems of estimating error probabilities without gold standard. In particular, we make three contributions.

First, we demonstrate that estimation of the error probabilities FAP and FRP is problematic since these are in general unidentifiable parameters in the model. This result goes a long way in explaining the poor performance of estimation methods found in simulation studies. The conclusion naturally carries over to similar concepts such as the probabilities of a false positive and a false negative, or the sensitivity and specificity.

Second, we show how an error probability can be decomposed into a systematic and a random component. The random error component in turn can be quantified by the IAP or IRP. These parameters can be estimated by modifying some existing methods, but Chapter 5 will propose techniques designed deliberately for this purpose. The IAP and IRP are useful results in themselves, as they indicate to what extent a test gives poorly reproducible (random) results, or at the other extreme, is capable of discriminating between worse and better items. Even a rough calibration of a test turns $\widehat{IAP}$ and $\widehat{IRP}$ into approximations for FAP and FRP.

Finally, we show by analysis that nuisance effects should be taken into account while designing the MSA study. Experimental randomization averages out their effects and thereby treats them as part of the random test errors *IAP* and *IRP*.

Chapter 4

Estimation of FAP and FRP without gold standard: simulation study

The previous chapter showed that estimation of the false acceptance and rejection probabilities (*FAP* and *FRP*) is inherently problematic in the absence of a gold standard. Keeping in mind George Box's maxim that "all models are wrong, some are useful", this chapter investigates whether existing methods give useful approximations of the *FAP* and *FRP*. The conclusion is negative: putting existing methods to the test, we show that estimation biases are often substantial. This is the last chapter that investigates the shortcomings of current methods, concluding the first goal of this thesis.

Regarding the second goal, which was to find alternative, reliable, quantifiers of measurement error, the previously introduced inconsistent acceptance and rejection probabilities (*IAP* and *IRP*) will prove to be far more reliably estimable.

4.1 Introduction

This chapter compares the main methods that are proposed in industrial statistics and engineering for the evaluation of binary tests without a gold standard, to learn how robust they are to model misspecification. In particular, we focus on evaluations

that yield estimates of misclassification probabilities:

$$FAP = P[Y = 1|X = 0] \text{ (False Acceptance Probability)},$$
$$FRP = P[Y = 0|X = 1] \text{ (False Rejection Probability)},$$

where $X = 1$ for conforming products and $X = 0$ for defective products. Moreover, $Y = 0$ represents rejection, and $Y = 1$ represents acceptance. We focus on situations where a gold standard is not available, i.e. the true value X is unknown.

Warnings have been voiced by Albert and Dodd (2004) and in Chapter 3 regarding the reliability of current methods for estimating FAP and FRP: they might not be reliable when a gold standard is unavailable. Nevertheless, we believe that a more solid exploration of the performance of existing models is neccessary. There are many situations not considered by Albert and Dodd (2004), and the conclusions in Chapter 3 are based on mathematical properties of the estimators, instead of empirical evidence. We propose a systematic comparison of various approaches. We evaluate the performance (in terms of estimator inconsistency) of a variety of methods in a variety of situations. In addition to situations that are given in literature, we also consider situations in which there are nuisance variables that influence outcomes of MSA experiments. Robustness to nuisance variables has not been thoroughly considered in literature yet.

The study is based on a crossed design where methods $M1, M2$ and $M3$ are applied in a range of scenarios $S1, S2, \ldots$. The estimation bias of the methods in the various scenarios is analyzed with three goals:

- Establish which method is most reliable across a range of realistic scenarios.
- Empirically establish how problematic estimation of FAP and FRP is without gold standard.
- If the problems turn out to be severe: identify alternative ways to evaluate binary tests.

The next section 4.2 explains the set-up of the study in detail. Section 4.3 presents the findings of the study, which results in specific conclusions discussed in the final section.

4.2 Theory and methods

The current literature describes three classes of approaches for estimating misclassification probabilities of binary tests under absence of a gold standard: traditional latent class methods ($M1$), latent class random effects approaches ($M2$), and mechanistic approaches based on characteristic curves ($M3$). In the comparison study we

try these methods $M1, M2$ and $M3$ in a number of scenarios $(S1, S2, \ldots)$ to learn how sensitive they are to violations of their assumptions.

In this section, we present the statistical modelling framework for the study, and we briefly describe the three methods and the scenarios in which they were tried. The descriptions of the methods are cursory, only highlighting the main idea, but full descriptions can be found in the references.

4.2.1 Statistical modeling

When a test is applied in regular production, it produces results $Y_i = 0$ (rejection) or $Y_i = 1$ (acceptance) for tested items $i = 1, 2, \ldots$. The unobservable true states of the items are $X_i = 0$ (truly defective) or $X_i = 1$ (conforming to specifications), and the (unknown) defect rate is $p = P[X_i = 0]$. The probability that an item i is rejected is $R_i = P[Y_i = 0]$, which depends on X_i and possibly on other properties affecting the measurement.

To determine the misclassification probabilities of the test, one executes an MSA experiment, in which a sample of I items are tested K times, producing results $Y_{ik} \in \{0, 1\}$. When a gold standard is available, the corresponding true values $X_1, \ldots, X_I$ are established, and a comparison of the $\{Y_{ik}\}_{k=1,\ldots,K}$ to X_i allows the calculation of FAP and FRP. In the problem under consideration, however, a gold standard is unavailable and the true states X_i of the items in the experiment are unobservable. In fact, one can only observe the per-item rejection counts $U_i = \sum_{k=1}^{K} (1 - Y_{ik}) \in \{0, 1, \ldots, J\}$, from which the rejection probabilities R_i can be estimated.

In this chapter, we will specify statistical models in terms of the R_i, avoiding the unobservable X_i. In the traditional latent class model, $R_i = 1 - FAP$ for all defective items and $R_i = FRP$ for all conforming items (with FAP and FRP two constants). Our model allows variation in the rejection probabilities of conforming and defective items, and we treat the R_i as a random effect. The distribution of rejection probabilities in the population of items is $F_R(r) = P[R_i \leq r]$. Further, we assume that given an item's rejection probability, repeated tests are independent: conditional on the event $\{R_i = r\}$, the $Y_{i1}, \ldots, Y_{iK}$ are i.i.d. Bernoulli (with parameter $1 - r$) and the rejection counts U_i have a binomial (K, r) distribution.

The distribution of rejection probabilities can be interpreted as a mixture of two component distributions: $F_R(r) = p \times F_R^0(r) + (1 - p) \times F_R^1(r)$, with $F_R^0(r) = P[R_i \leq r | X_i = 0]$ (the distribution of rejection probabilities of defective items) and $F_R^1(r) = P[R_i \leq r | X_i = 1]$ (the distribution of rejection probabilities of conforming items).

Figure 4.1 gives an example. The solid curve is the density f_R of rejection probabilities in the population of items. The gray and white areas show the components

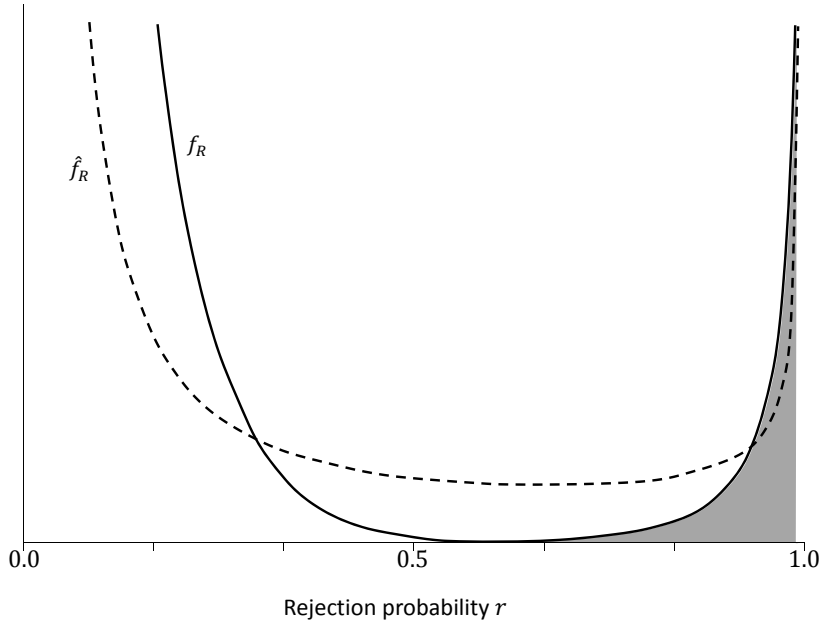


Figure 4.1: Density f_R of rejection probabilities in scenario $S2a$ (solid curve). Estimated pdf ($\widehat{f_R}$) fitted by method $M3$ (dashed curve).

$p \times f_R^0$ and $(1-p) \times f_R^1$ associated with defective and conforming items. In this example (produced by scenario $S2a$ explained later), the rejection probabilities of conforming and defective items are clearly separated and $f_R(r) \approx 0$ around $r = 0.6$.

We define the FAP and FRP as the mean probabilities of a misclassification, that is:

$$FAP = E[1 - R_i | X_i = 0],$$
$$FRP = E[R_i | X_i = 1].$$

In Chapter 3 it has also been suggested to evaluate a binary test in terms of two related metrics, the probabilities of inconsistent rejection and acceptance:

$$IAP = E[1 - R_i | X_i = 0] = P\left[Y_{ik} = 1 | \tilde{X}_i = 0\right],$$
$$IRP = E[R_i | X_i = 1] = P\left[Y_{ik} = 0 | \tilde{X}_i = 1\right].$$

Here, $\tilde{X}_i$ is the modal (most likely) test result for item i, that is, $\tilde{X}_i = 0$ if $R_i > 0.5$ and $\tilde{X}_i = 1$ if $R_i \leq 0.5$. The previous chapter shows that the errors FAP and FRP can be decomposed into systematic and random error components, and that the latter can be expressed as the IAP and the IRP. They represent the probability that a test result randomly deviates from the most likely result for the item. They are useful measures of the reproducibility of the binary tests, analogous to the precision of numerical measurements. However, IAP and IRP do not reflect the systematic error, i.e., the extent to which the most likely outcome is the true state.

4.2.2 Methods included in the comparison

In the study we compare the sensitivity to model (mis)specification of three methods for estimating the FAP and FRP, and/or the IAP and IRP, proposed in the recent literature in industrial statistics.

M1: Traditional latent class method The most basic approach assumes that FAP and FRP are fixed effects. Methods such as those in Hui and Walter (1980); Boyles (2001); Van Wieringen and De Mast (2008); Danila et al. (2010) and Beavers et al. (2011) fit latent class models in which FAP and FRP are constants in the subpopulations of conforming and defective items. Method $M1$ in our comparison fits a model with $R_i = 1 - FAP$ if $X_i = 0$ and $R_i = FRP$ if $X_i = 1$. The model has three parameters: FAP, FRP and $p = P[X_i = 0]$. Under the simple setting of $M1$ we have, under the mild assumption that $FAP < 0.5$ and $FRP < 0.5$, that $IAP = FAP$ and $IRP = FRP$. In view of this, the estimated IAP and IRP are identical to the estimated FAP and FRP.

M2: Latent class random effects approach An extension of the traditional latent class method allows FAP and FRP to be random effects. Danila et al. (2012) propose a method based on a mixture of two beta distributions: f_R^0 is a beta distribution with mean $1-\mu_0 = 1-FAP$ and dispersion ϕ_0, and f_R^1 is a beta distribution with mean $\mu_1 = FRP$ and dispersion ϕ_1. The model has five parameters: $\mu_0, \mu_1, \phi_0, \phi_1$ and p. Based on the beta distributions in each subpopulation, the density of the rejection probabilities in the entire items population is the mixture

$$f_R(r) = (1-p) \times \frac{r^{\mu_1 \frac{1-\phi_1}{\phi_1}} (1-r)^{(1-\mu_1)\frac{1-\phi_1}{\phi_1}-1}}{\text{Beta}\left(\mu_1 \frac{1-\phi_1}{\phi_1}, (1-\mu_1)\frac{1-\phi_1}{\phi_1}\right)} + p \times \frac{r^{(1-\mu_0)\frac{1-\phi_0}{\phi_0}-1} (1-r)^{\mu_0 \frac{1-\phi_0}{\phi_0}-1}}{\text{Beta}\left((1-\mu_0)\frac{1-\phi_0}{\phi_0}, \mu_0 \frac{1-\phi_0}{\phi_0}\right)}.$$

FAP and *FRP* are estimated from the fitted μ_0 and μ_1. As introduced originally, the method does not produce estimates for *IAP* and *IRP*. However, estimates for *IAP* and *IRP* can be found from the fitted density $\widehat{f}_R$ by

$$\widehat{IAP} = \frac{\int_{0.5}^{1} (1-r)\, \widehat{f}_R(r)\, \mathrm{d}r}{\int_{0.5}^{1} \widehat{f}_R(r)\, \mathrm{d}r},$$

$$\widehat{IRP} = \frac{\int_{0}^{0.5} r \widehat{f}_R(r)\, \mathrm{d}r}{\int_{0}^{0.5} \widehat{f}_R(r)\, \mathrm{d}r}.$$

M3: Approaches based on characteristic curves The approach in De Mast et al. (2011) and Erdmann et al. (2016a) models the source of variability among items more explicitly. It assumes that the dichotomous true state masks an underlying continuum where some items are more defective than others (and, possibly, some items are more conforming than others). Thus, the true state X_i is determined by a latent continuous quality characteristic $Z_i \in \mathbb{R}$ in the sense that $X_i = 1$ if $Z_i \leq USL$ and $X_i = 0$ otherwise (where *USL* is the implicitly defined upper specification limit on the unobservable quality characteristic Z_i).

Variability in the rejection probabilities, then, is related to variability in Z_i by means of a characteristic curve (here a logistic curve with inflection point at $z = \delta$ and slope parameter α)

$$q(z) = P[Y_{ik} = 0 | Z_i = z] = \left(1 + \exp\{-\alpha(z-\delta)\}\right)^{-1}.$$

Writing $F_Z(z) = P[Z_i \leq z]$ for the distribution of Z_i in the population of items, the distribution of rejection probabilities is

$$F_R(r) = P[q(Z_i) \leq r] = F_Z\left(q^{-1}(r)\right).$$

Erdmann et al. (2016a) take F_Z to be the standard normal distribution. The model has two parameters: α and δ. From the fitted characteristic curve $\widehat{q}(z)$, *IAP* and *IRP*

are determined from:

$$\widehat{IAP} = \int_{\widehat{\delta}}^{\infty} \left(1 - \widehat{q}\left(z\right)\right) \phi\left(z\right) \mathrm{d}z \Big/ \int_{\widehat{\delta}}^{\infty} \phi\left(z\right) \mathrm{d}z,$$

$$\widehat{IRP} = \int_{-\infty}^{\widehat{\delta}} \widehat{q}\left(z\right) \phi\left(z\right) \mathrm{d}z \Big/ \int_{-\infty}^{\widehat{\delta}} \phi\left(z\right) \mathrm{d}z.$$

The approach does not offer a way to estimate FAP and FRP except in the (unrealistic) case that $q\left(USL\right)$ is known.

For all three methods, recent literature has investigated the suitability of various sampling strategies. For the MSA study, one needs a sample $i = 1, \ldots, I$ of items. The most straightforward sampling strategy is to take a representative sample from the total population of relevant items. However, given the usually very low defect rates of industrial processes, samples thus collected typically contain no or only very few items with $R_i > 0.5$. Recent studies have investigated the effect of various alternative sampling strategies on the precision of the estimates (Danila et al., 2010, for approaches of type $M1$; Danila et al., 2012 for $M2$; and Erdmann et al., 2016a for $M3$). Consistently, these studies show that the best sampling strategy is to take a representative sample from the stream of rejected items, which typically produces a sample with more evenly spread R_i (note that under realistic FAP, FRP and p even the stream of rejected items contains more good items, rejected falsely, than defective items; see De Mast et al., 2011). However, such a sample is not representative for the total population of items. As parameters are estimated by maximum likelihood, it is possible to correct for the resulting bias in the estimators, by calculating the likelihood contributions conditional on the event that they have been rejected initially (see Section 4.2.5).

A further improvement can be obtained by augmenting the data from the MSA experiment with an estimate of the rejection rate $P\left[Y_i = 0\right]$ based on historical data. Even if such historical data are not available, they can be recorded as a by-product in the time period when the sample is collected from the stream of rejected items. Recent papers (Danila et al., 2010, 2012; Erdmann et al., 2016a) call such data baseline data and demonstrate how their incorporation in the estimation procedure makes estimates more precise. It is this setup, where the MSA study involves a sample of I items taken from the rejection stream plus a historical estimate for the rejection rate, that we have implemented for all methods $M1, M2$ and $M3$.

4.2.3 Scenarios

To analyze the sensitivity of methods $M1, M2$ and $M3$ to model misspecification, their estimation bias was compared in 10 scenarios. Scenario $S1$ is a situation where measurements result from the simple latent class model (as in method $M1$). In scenarios $S2a, S2b$ and $S2c$, data result from random effects models (as in $M2$), in which the R_i have beta ($S2a, b$) or logit-logistic ($S2c$) distributions with little ($S2a$) or substantial ($S2b, c$) dispersion. Scenarios $S3a$ through $S3d$ (related to $M3$) are based on a standard normal continuous characteristic Z. This characteristic determines the rejection probability through a steep logistic characteristic curve in $S3a$, and a relatively flat logistic curve in $S3b$. In $S3c$ the characteristic curve is an asymmetric, cumulative Weibull distribution, and in $S3d$ it is a cumulative Weibull distribution with an infimum set at $\inf_z q(z) = 0.01$ instead of 0.00.

Scenarios $S4a$ and $S4b$ have characteristic curves that incorporate the effects of nuisance variables. In real life, these could be the effects of conditions under which the tests are performed (such as light conditions for visual inspections) or characteristics of the appraisers (such as fatigue). In the study, such effects are modelled as a standard normal random variable V_{ik}, which may assume a new value for each item i and even each repeat k because repeats are usually dispersed over time to minimize memory effects. In scenario $S4a$, it affects the characteristic curve's slope in its inflection point:

$$P[Y_{ik} = 0 | Z_i = z, V_{ik} = v] = q(z, v) = (1 + \exp\{-10e^v (z - 2)\})^{-1}$$

and in $S4b$ it affects the location of the curve's inflection point:

$$P[Y_{ik} = 0 | Z_i = z, V_{ik} = v] = q(z, v) = (1 + \exp\{-10 (z + v - 2)\})^{-1}$$

Details of the scenarios can be found in Section 4.2.5.

4.2.4 Implementation details

In our study, the parameters of the models prescribed by methods $M1$ (p, FAP and FRP), $M2$ (p, μ_0, μ_1, ϕ_0 and ϕ_1) and $M3$ (α and δ) are estimated by maximum likelihood, as is often used for fitting these models (Van Wieringen and De Mast, 2008; Danila et al., 2012; Erdmann et al., 2016a and many others). We implemented the recommended setup consisting of I items sampled from the stream of rejected items, which are tested K times, and in addition, I^{His} baseline data are included in the likelihood function.

As the study focusses on robustness against model misspecification rather than the determination of appropriate sample sizes, the comparison study treats sample sizes as neutrally as possible. Recommendations in the literature for the number of

repeats are in the range of $K = 7$ (Erdmann et al., 2016a for method $M3$) to $K = 10$ (Danila et al., 2012 for $M2$), with $K = 3$ as a minimum to ensure identifiability (Van Wieringen and De Mast, 2008 for $M1$). In our study, we have worked with $K = 7$. The size of the sample of items is left out of consideration by studying the limit behavior of estimators when I approaches infinity. For this reason, estimators converge to the true values when the model assumed in a scenario is the same as the model assumed in the method (such as method $M1$ applied in scenario $S1$). The number I^{His} of historical data is taken in a fixed proportion to the sample size I, namely, $I^{\text{His}} = I/P[Y_{ik} = 0])$. This is the ratio that one finds if no historical data are available, and instead, the baseline data are collected as a by-product of obtaining the sample of I rejected items. Namely, in order to obtain I rejected items, one has to produce an expected number of $I/P[Y_{ik} = 0])$ items. When historical data are available, the number I^{His} of baseline data will almost always be larger.

Likelihood functions and other details of the implementation are given in Section 4.2.5. For our calculations, we used Mathematica 8 (Wolfram Research, 2010). Log-likelihood functions are optimized with the interior point algorithm (as implemented in the function"FindMaximum") or the Nelder-Mead algorithm (as implemented in "NMaximize"). The integrals are approximated numerically using adaptive quadrature (as implemented in "NIntegrate").

4.2.5 Simulation details

Each of the methods $M1$, $M2$ and $M3$ fits a model by maximum likelihood estimation on the basis of a sample of initially rejected items augmented with baseline data. We discuss here the details of this estimation procedure.

Data are generated by the models defined in each of the scenarios $S1-4$, and probabilities calculated under these data generating models are denoted P_{Sc} . The models fitted to the data depend on the methods $M1-3$, and probabilities calculated under these fitted models are denoted $P_{\boldsymbol{\theta}}$, with $\boldsymbol{\theta}$ the parameter vector of the fitted model. The results of the MSA study are the rejection counts $U_i = \sum_{k=0}^{K} (1 - Y_{ik})$, which can be aggregated in the response pattern frequencies $E_k = \{\#i|U_i = k\}$ for $k = 0, \ldots, K$ (with realizations e_k). For the I^{His} baseline data we have $E_0^{\text{His}} = \left\{\#i \in \left\{1, 2, \ldots, I^{\text{His}}\right\} : Y_i = 0\right\}$ as the number of rejected items, and E_1^{His} for the number of accepted items (and $E_0^{\text{His}}/I^{\text{His}}$ is the historical rejection rate). The log-likelihood of the data is

$$L(\boldsymbol{\theta}|\boldsymbol{E} = \boldsymbol{e}) = \sum_{k=0}^{K} e_k \log\{P_{\boldsymbol{\theta},\text{Rej}}[U_i = k]\} + \sum_{k=0}^{1} e_k^{\text{His}} \log\{P_{\boldsymbol{\theta}}[Y_i = k]\}. \tag{4.1}$$

Here,

$$P_{\boldsymbol{\theta}}\left[Y_i = 0\right] = \int_0^1 r f_R\left(r|\boldsymbol{\theta}\right) \mathrm{d}r;$$
$$P_{\boldsymbol{\theta}}\left[Y_i = 1\right] = 1 - P_{\boldsymbol{\theta}}\left[Y_i = 0\right],$$

with f_R as implied by the model underlying $M1, M2$ or $M3$. Further, since the I items in the sample are taken from the stream of rejected items, the likelihood is calculated conditional on this initial rejection:

$$P_{\boldsymbol{\theta},\mathrm{Rej}}\left[U_i = k\right] = P_{\boldsymbol{\theta}}\left[U_i = k|\text{item initially rejected}\right] = \frac{\int_0^1 f_R\left(r|\boldsymbol{\theta}\right)\binom{K}{k} r^{k+1}\left(1-r\right)^{K-k}\mathrm{d}r}{\int_0^1 f_R\left(r|\boldsymbol{\theta}\right) r \mathrm{d}r}.$$

The aim of the comparison study is to learn what can be estimated with a sufficiently large sample size I. We take sample size (that is, sampling error) out of consideration by studying what can be estimated if the size of the sample of items approaches infinity ($I \to \infty$). Also, we fix the ratio of I and the number I^{His} of baseline data at $I/I^{\mathrm{His}} = P_{\mathrm{Sc}}\left[Y_i = 0\right]$. We believe that often, many more baseline data are available, depending on how long the test system has been in use and whether results are logged. But a lower bound is obtained when the baseline data are collected as a by-product of obtaining the sample for the MSA study. Namely, in order to obtain I rejected items, the expected total number of items to be inspected is $I/P_{\mathrm{Sc}}\left[Y_i = 0\right]$, and these can be used as baseline data. Dividing the loglikelihood by I and substituting this ratio we obtain

$$\frac{L\left(\boldsymbol{\theta}|\boldsymbol{E}=\boldsymbol{e}\right)}{I} = \sum_{k=0}^{K} \frac{e_k}{I} \log P_{\boldsymbol{\theta},\mathrm{Rej}}\left[U_i = k\right] + \frac{1}{P_{\mathrm{Sc}}\left[Y_i = 0\right]} \sum_{k=0}^{1} \frac{e_k^{\mathrm{His}}}{I^{\mathrm{His}}} \log P_{\boldsymbol{\theta}}\left[Y_i = k\right].$$

If $I \to \infty$, the response pattern frequencies converge (with probability 1) to their expected values:

$$E_k/I \overset{a.s.}{\to} P_{\mathrm{Sc}}\left[U_i = k\right],$$

and also $I^{\mathrm{His}} \to \infty$ and $E_k^{\mathrm{His}}/I^{\mathrm{His}} \overset{a.s.}{\to} P_{\mathrm{Sc}}\left[Y_i = k\right]$. The maximum of $L\left(\boldsymbol{\theta}|\boldsymbol{E}=\boldsymbol{e}\right)$ is in the same location as that of $L\left(\boldsymbol{\theta}|\boldsymbol{E}=\boldsymbol{e}\right)/I$. Therefore, for large sample sizes ($I \to \infty$), the parameter values found from maximizing $L\left(\boldsymbol{\theta}|\boldsymbol{E}=\boldsymbol{e}\right)$ converge (with

probability 1) to

$$\arg\max_{\boldsymbol{\theta}} \sum_{k=0}^{K} P_{\mathrm{Sc}}\left[U_i = k\right] \times \log\left\{P_{\boldsymbol{\theta},\mathrm{Rej}}\left[U_i = k\right]\right\} + \frac{1}{P_{\mathrm{Sc}}\left[Y_i = 0\right]} \sum_{k=0}^{1} P_{\mathrm{Sc}}\left[Y_i = k\right] \times \log\left\{P_{\boldsymbol{\theta}}\left[Y_i = k\right]\right\},$$

which is the optimization problem we solve to find parameter estimates. To summarize, we leave the question of appropriate sample sizes out of consideration by studying the goodness of fit of methods $M1$, $M2$ and $M3$ when the sample size I becomes large. The estimated model parameters $\widehat{\boldsymbol{\theta}}$ yield a fitted density $\widehat{f_R}$, from which estimates for IAP, IRP, and possibly FAP and FRP, are derived.

4.3 Findings

Applying the various methods across the selected scenarios reveals to what extent they are robust against violations of the assumptions on which they are based. Table 4.1 presents the results. We discuss our findings for the methods $M1$, $M2$ and $M3$.

4.3.1 Traditional latent class method

Our findings for the performance of the traditional latent class method M1 confirm earlier findings in the medical statistics literature (Torrance-Rynard and Walter, 1997; Albert and Dodd, 2004). Namely, such methods only work when the rejection probabilities are constant in the subpopulations of conforming and defective items (scenario $S1$), but seriously derail in other scenarios, where M1 gives badly biased estimates for FRP, FAP, IRP and IAP. Figure 4.2 shows a typical example. Here, the solid line is the density f_R of rejection probabilities produced by scenario $S2b$, where the rejection probabilities of conforming and defective items follow beta distributions. The implied misclassification probabilities are $FAP = 0.05$ and $FRP = 0.05$. Further, $IAP = 0.1293$ and $IRP = 0.0255$. Method $M1$ yields $\widehat{p} = 0.0882$, $\widehat{FAP} = \widehat{IAP} = 0.1220$ and $\widehat{FRP} = \widehat{IRP} = 0.0357$ (which gives point masses $P\left[R_i = 0.0357\right] = 0.9118$ and $P\left[R_i = 0.878\right] = 0.0882$). Note that $\widehat{FAP}$ and $\widehat{FRP}$ are both strongly biased. The estimates for IAP and IRP are more accurate, but in other scenarios ($S3a$, $S3b$, $S3c$) the biases are enormous. In this and similar situations, $M1$ approximates a continuous density f_R of rejection probabilities by two point masses, and the resulting estimates are unreliable.

Scenario	Method	FAP	FRP	IAP	IRP
$S1$	$TRUE$	0.0500	0.0500	0.0500	0.0500
Constant FRP	$M1*$	0.0500	0.0500	0.0500	0.0500
and FAP	$M2$	0.0500	0.0500	0.0500	0.0500
	$M3$	NA	NA	0.2153	0.0358
$S2a$	$TRUE$	0.0500	0.0500	0.0540	0.0497
Beta distributions	$M1$	0.0535	0.0692	0.0535	0.0692
(little dispersion)	$M2*$	0.0500	0.0500	0.0540	0.0497
	$M3$	NA	NA	0.2001	0.0330
$S2b$	$TRUE$	0.0500	0.0500	0.1293	0.0255
Beta distributions	$M1$	0.1220	0.0357	0.1220	0.0357
(substantial dispersion)	$M2*$	0.0500	0.0500	0.1293	0.0255
	$M3$	NA	NA	0.1359	0.0200
$S2c$	$TRUE$	0.0500	0.0500	0.0995	0.0405
Logit-logistic	$M2$	0.1891	0.0373	0.0970	0.0399
distributions	$M3$	NA	NA	0.1765	0.0280
$S3a$	$TRUE$	0.5360	0.0000	0.0733	0.0022
Std. normal f_Z	$M1$	0.0206	0.4179	0.0206	0.4179
steep logistic	$M2$	0.1598	0.0004	0.0799	0.0014
characteristic curve	$M3*$	NA	NA	0.0733	0.0022
$S3b$	$TRUE$	0.0252	0.0183	0.2863	0.0278
Std. normal f_Z	$M1$	0.2673	0.1584	0.2673	0.1584
flat logistic	$M2$	0.5614	0.0205	0.2846	0.0275
characteristic curve	$M3*$	NA	NA	0.2863	0.0278
$S3c$	$TRUE$	0.5587	0.0000	0.0538	0.0005
Std. normal f_Z	$M1$	0.0135	0.5895	0.0135	0.5895
cum. Weibull	$M2$	0.1030	0.0000	0.0540	0.0005
characteristic curve	$M3$	NA	NA	0.0476	0.0012
$S3d$	$TRUE$	0.5531	0.0100	0.0538	0.0105
Std. normal f_Z	$M1$	0.0370	0.0284	0.0370	0.0284
cum. Weibull	$M2$	0.1040	0.0100	0.0540	0.0105
characteristic curve	$M3$	NA	NA	0.1439	0.0071
$S4a$	$TRUE$	NA	NA	0.1522	0.0137
Std. normal f_Z, f_V	$M1**$	0.2217	0.0120	0.2217	0.0120
logistic characteristic	$M1$	0.1628	0.0157	0.1628	0.0157
curve, threshold	$M2**$	0.1843	0.0161	0.1534	0.0137
influenced by	$M2$	0.2231	0.0120	0.1689	0.0132
nuisance V	$M3**$	NA	NA	0.1894	0.0111
	$M3$	NA	NA	0.2142	0.0133
$S4b$	$TRUE$	NA	NA	0.3659	0.0675
Std. normal f_Z, f_V	$M1**$	0.0731	0.0125	0.0731	0.0125
logistic characteristic	$M1$	0.5468	0.0440	0.5468	0.0440
curve, steepness	$M2**$	0.0744	0.0153	0.0809	0.0088
influenced by	$M2$	0.7173	0.0520	0.3650	0.0672
nuisance V	$M3**$	NA	NA	0.0812	0.0088
	$M3$	NA	NA	0.3547	0.0678

Table 4.1: Overview of the results when methods $M1-3$ are applied in scenarios $S1-4$.
*: Method fits the exact same model as in the scenario, and therefore estimates converge to the true parameter values (for $I \to \infty$).
**: Method applied without proper experimental randomization.
NA: The $M3$ methods do not offer a way to estimate FAP and FRP.

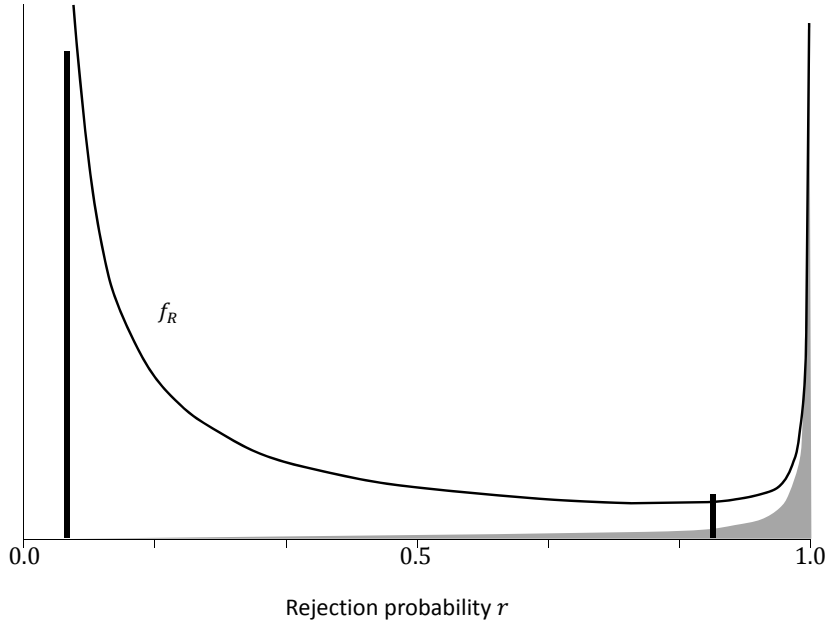


Figure 4.2: True pdf f_R of rejection probabilities in scenario $S2b$ (solid curve), decomposed into the densities for conforming (white) and defective (gray) items. Vertical bars represent the distribution of rejection probabilities fitted by method $M1$.

4.3.2 Approaches based on characteristic curves

The method $M3$ based on a characteristic curve model produces reliable estimates for IAP and IRP under scenarios $S3a - b$, which are instances of its native model. However, the method turns out to lack robustness in many other scenarios and produces a poor fit. The performance is particularly poor in scenarios where there is little variation in the rejection rates R_i of the items ($S1, S2a, S2c, S3c$ and $S3d$). We illustrate the problem with the typical example of scenario $S2a$ (Figure 4.1). The solid curve is f_R with $FAP = FRP = 0.0500, IAP = 0.0540$ and $IRP = 0.0497$. There is little variation in rejection probabilities, which can be seen from f_R being almost zero for rejection probabilities between 0.4 and 0.8. The characteristic curve approach $M3$ turns out to be unable to fit this low degree of variation ($M3$ produces the dashed curve with $\widehat{\alpha} = 4.56$ and $\widehat{\delta} = 1.43$). $M3$ does not offer a way of estimating FAP and FRP, but instead, the method yields the estimates $\widehat{IAP} = 0.2001$ and $\widehat{IRP} = 0.0330$, which are rather far off.

4.3.3 Latent class random effects method

The latent class random effects approach $M2$ gives a good fit of f_R across scenarios, which is a promising finding. The $M2$ method fits a model with 5 parameters, 3 more than $M3$, which may explain the relatively good performance. For many scenarios, however, the estimated FAP and FRP are substantially off. To illustrate, the true f_R in scenario $S2c$ is a mixture of logit-logistic distributions, but it is approximated rather well by the mixture of beta distributions fitted by method $M2$. However, the estimated FAP and FRP (0.1891 and 0.0373) are far off (true values 0.0500 and 0.0500). The estimated FAP and FRP are further off in scenarios where the component densities f_R^0 and f_R^1 have more dispersion, and therefore, are less clearly separated.

Estimation of FAP and FRP becomes hopeless for the $S3$ scenarios, where the true state X_i is a dichotomization of a continuous characteristic Z_i and the two component densities f_R^0 and f_R^1 are less clearly separated. For example, in Figure 4.3 (scenario $S3b$), the density f_R of all items is fitted rather well by M2. The component densities f_R^1 for conforming items (that is, $Z_i \leq USL$) and f_R^0 for defective items ($Z_i > USL$) are separated by a vertical bound at $r = q\,(USL) = 0.256$, but this is impossible to determine from the shape of f_R alone. Consequently, the estimated FAP and FRP are far off (true values: 0.0252 and 0.0183; estimates: 0.5614 and 0.0205).

Still, because of the good fit of f_R, method $M2$ produces reliable estimates for IAP and IRP across the scenarios $S1 - S3$. For example, in scenario $S3b$ (Figure 4.3) $\widehat{IAP} = 0.2846$ and $\widehat{IRP} = 0.0275$ (true values are 0.2863 and 0.0278). In scenario $S2c$, the procedure gives 0.0970 and 0.0399 (true values 0.0995 and 0.0405).

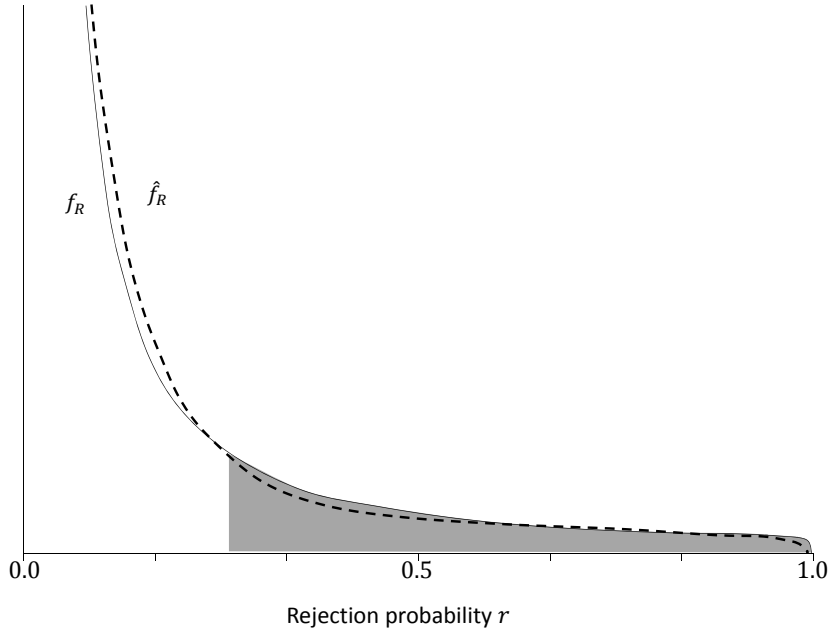


Figure 4.3: True pdf f_R of rejection probabilities in scenario $S3b$ (solid curve), with the densities of conforming and defective items marked by white and gray. Estimated pdf $\widehat{f}_R$ fitted by method $M2$ (dashed curve).

4.3.4 Estimation in the presence of nuisance variables

The scenarios $S4a - b$ explicitly incorporate the effects of a nuisance variable V_{ik} on the test results. Following Akkerhuis et al. (2016b) and Chapter 3, experimental randomization should be able to deal with nuisance variables effectively. This means that each of the K repeats per item are done under a new realization of the nuisance variables, and such that these realizations are representative for their population distribution. Experimental randomization was implemented in our study by drawing a new realization for V_{ik} from its assumed population distribution F_V for each repeat k and each item i. The best results are obtained from method $M2$. In $S4b$, for instance, it gives $\widehat{IAP} = 0.3650$ and $\widehat{IRP} = 0.0672$ (true values 0.3659 and 0.0675).

The importance of experimental randomization is illustrated by applying the methods under a more careless setup (marked by a double star in Table 4.1), where only a single realization of V_{ik} is drawn per item i. Under this arrangement, $M2$ gives $\widehat{IAP} = 0.0809$ and $\widehat{IRP} = 0.0088$.

4.4 Discussion and conclusions

Latent variable methods are an obvious option to deal with gold standard unavailable situations. Literature has explored various methods to do so, but there is no consensus about their performance. And in fact, the mathematical analysis in the previous chapter shows that generally, FAP and FRP are unidentifiable when a gold standard is unavailable. In this study we have investigated when and whether currently available methods are suitable, and if not, how severe the problems are. The first set of conclusions mirrors the results of a similar study for methods proposed in medical statistics (Albert and Dodd, 2004): all methods work well if their model assumptions are precisely met, but estimates for FAP and FRP easily become severely biased otherwise. Especially under the (realistic) scenarios $S3$, estimation of FAP and FRP becomes hopeless. Taken together with the mathematical analysis in the previous chapter, these results constitute a strong case that the estimation of the FAP and FRP of a binary test without gold standard is in general an unattainable ambition. Especially if one is not satisfied with evaluating a binary test in terms of IAP and IRP, this is a strong motivation to go quite some way in finding a gold standard.

By analogy with MSA studies for numerical measurements, the previous chapter shows how the total misclassification errors FAP and FRP can be decomposed into systematic errors and random errors. The latter can be quantified as the IAP and IRP. Our current study shows that IAP and IRP can reliably be estimated by a modification of method $M2$, and that even the presence of nuisance effects can be handled by carefully randomizing repeated tests. If IAP and IRP are poor, this means that there is too much random variability in the test system's results to be informative,

and even a good calibration will not save such system. If IAP and IRP are good, this means that the test systems discriminative capability is sufficient for discerning worse from better items: the test results are reproducible. For many applications, this will in itself be a useful result. If good reproducibility is not enough, and an application demands that also the systematic error is controlled, the system needs a calibration. This is problematic without a gold standard. For test systems with good IAP and IRP, a rough calibration will typically do, and it may be possible to do so on the basis of a training set of items that are rated as defective or conforming by consensus among experts.

The model fitted in $M2$ is a mixture of two beta distributions, reflecting the original purpose of estimating FAP and FRP by fitting probability distributions for conforming and nonconforming items separately. For the purpose of evaluating a test in terms of its IAP and IRP, the class of distributions that is fitted to the data does not need to be a mixture of two distributions corresponding to conforming and defective items. This leaves a wide class of candidate distribution functions open for exploration, with the aim of finding a class of functions that has a limited number of parameters, but that nevertheless can reliably fit a wide range of scenarios. The next chapter elaborates a method that fits a model based on a squared polynomial multiplied by a beta distribution, whose order depends on the fit to the data. It also presents well-grounded recommendations for sample size and study design, as well as diagnostic and goodness-of-fit tests to complement the analysis.

All in all, the study offers bad and good news: without gold standard, the estimation of the error probabilities FAP and FRP is usually not possible, but it is possible to determine the random components IAP and IRP of these errors. Although this conclusion is in line with theory and practice in MSA studies for numerical measurements, it represents quite a drastic turn in the development of techniques for binary MSA.

Chapter 5

Robust estimation of IAP and IRP

Previous chapters have shown that estimation of the false acceptance and rejection probabilities (*FAP* and *FRP*) in the absence of a gold standard is problematic, and confirmed that none of the models that are designed to do so are reliable unless their assumptions are *exactly* met. However, previous chapters found that these probabilities have random components, and that these are in fact estimable.

The second goal of this thesis is to find alternative and reliable quantifiers, and although these quantifiers have already been introduced, a reliable method of estimating them is lacking. This chapter proposes a robust method for estimating the random components of binary measurement error, as expressed by inconsistent acceptance and rejection probabilities (*IAP* and *IRP*).

5.1 Introduction

Earlier chapters introduced the inconsistent acceptance and rejection probabilities (*IAP* and *IRP*). The *IAP* and *IRP* are defined similarly as *FAP* and *FRP*, but condition on a different event:

$$IAP = P\left[Y = 1|\tilde{X} = 0\right],$$
$$IRP = P\left[Y = 0|\tilde{X} = 1\right],$$

with $\tilde{X}$ the most likely (modal) measurement outcome. Chapter 3 showed that these quantify the random components of the total measurement errors *FAP* and *FRP*.

While *FAP* and *FRP* compare a measurement outcome to the correct value X (hence the word false in *FAP* and *FRP*), *IAP* and *IRP* compare to the value that the test consistently tends to assign to the part (with probability of more than 0.5; hence the word inconsistent). Chapters 3 and 4 show that these random components are generally all one can hope to estimate when a gold standard is unavailable.

In view of the often critical ramifications of errors of binary tests, reliable methods for binary MSA are crucial. All of the current methods fail to deal effectively with the common complication of the unavailability of a gold standard, and the aspiration of estimating *FAP* and *FRP* in particular has been shown to be unattainable. Within the limitations imposed by the unavailability of a gold standard, this chapter proposes an efficient and robust method for estimation of the random errors *IAP* and *IRP*. The method is designed to not depend on hard-to-verify assumptions. We assess the method's robustness under a wide range of scenarios, and give recommendations for suitable sample sizes.

The second section will go into the statistical modeling and give a formal definition of *IAP* and *IRP*. The third section explains the estimation procedure, with the fourth section studying the methods robustness and offering sample size recommendations. Section 5 describes the application of the proposed method to a real-life example. Section 6 reviews the contribution of the new approach.

5.2 Statistical modeling

In a typical MSA study in a gold standard unavailable setting, a sample of parts $i = 1, \ldots, I$ is measured K times, giving $I \times K$ measurement outcomes $Y_{ik} \in \{0, 1\}$. We summarize the K repeated measurements of part i by the number of rejections $S_i = K - \sum_{k=1}^{K} Y_{ik}$, also called "response pattern". In the traditional latent class approach, the probability that a part i is rejected is $P[Y_{ik} = 0] = 1 - FAP$ for each nonconforming part, and $P[Y_{ik} = 0] = FRP$ for each conforming part (with *FAP* and *FRP* two constants). However, as mentioned in earlier chapters, rejection probabilies of conforming and nonconforming parts need not be constants, due to the effects of test conditions and nuisances, and also because the true state may be related to a multidimensional or continuous property. Therefore, we allow the rejection probabilities to vary: the rejection probability R_i of part i is a random effect drawn from a distribution $F_R(r) = P[R_i \leq r]$ with $r \in [0, 1]$. We assume that, conditional on R_i, the $Y_{i1}, \ldots, Y_{iK}$ are independent Bernoulli $(1 - R_i)$ variables, and S_i has a binomial (R_i, K) distribution.

As for the set-up of the MSA experiment, we build on the insights from recent literature. The basic set-up is to collect a sample $i = 1, \ldots, I$ of items that are measured K times. We follow the recommendation in previous literature (Danila et al.,

2010, 2012; Erdmann, 2012; Akkerhuis, 2012) to take a sample from the stream of rejected parts. This stream has a more balanced mix of nonconforming (rejected justly) and conforming (rejected falsely) parts, and in particular more evenly spread R_i (Erdmann, 2012). The more obvious choice of a sample from the total population of parts, to the contrary, usually contains almost no parts with $R_i \leq 0.5$, given the typically very low reject rates of industrial processes. This would result in a poor precision for the estimated IAP in particular. Our strategy to sample from the subpopulation of parts that initially have been rejected is called conditional sampling in Danila et al. (2010). The distribution F_R^{Rej} of rejection probabilities in the subpopulation of rejected parts is related to the distribution F_R^{Tot} in the total parts population as follows:

$$F_R^{\text{Tot}}(r) = P\left[R_i \leq r | \text{part initially rejected}\right] = \frac{P\left[R_i \leq r, \text{rejected}\right]}{P\left[\text{rejected}\right]} = \frac{\int_0^r t f_R^{\text{Tot}}(t)\,\mathrm{d}t}{\int_0^1 t f_R^{\text{Tot}}(t)\,\mathrm{d}t}.$$

Assuming that the test under study has been in use for some time in the relevant population of parts, we follow the recommendation of augmenting the sample with "baseline data". These are the number of rejections during a run of I^{Base} single inspections while the test system is in use in the regular production process. Danila et al. (2010, 2012); Erdmann (2012) and Akkerhuis (2012) show (for their methods) how the incorporation of this historical rejection rate improves the precision of estimates. These parts have rejection probabilities R_i drawn from F_R^{Tot}. Even if such baseline data are not available, they often can be collected as a by-product while gathering the sample from the stream of rejected parts: in order to obtain I rejected parts, on average a total of $I/E\left[R_i\right]$ parts need to be produced and inspected. By recording the number I^{Base} of parts inspected until I rejected parts have been collected, the baseline consists of I rejected parts and $I^{\text{Base}} - I$ accepted parts. Although in some situations more baseline data may be available, we assume that this is not the case in this chapter. The additional advantage of collecting baseline data in this manner is that all measurements are done in the same period, minimizing influence of time-dependent nuisance variables.

The experimenter ends up with I parts measured K times giving response patterns $S_i \in \{0, \ldots, K\}, i = 1, \ldots, I$ and I^{Base} parts measured once ($S_i^{\text{Base}} \in \{0, 1\}, i =$

$1, \ldots, I^{\text{Base}}$). These data lead to the following likelihoods L and L^{Base}:

$$L = \prod_{i=1}^{I} P\left[S_i = s_i\right] = \prod_{i=1}^{I} \int_0^1 P\left[S_i = s_i | R_i = t\right] f_R^{\text{Rej}}(t)\,\mathrm{d}t \propto$$

$$\prod_{i=1}^{I} \int_0^1 t^{s_i} (1-t)^{K-s_i} f_R^{\text{Rej}}(t)\,\mathrm{d}t,$$

$$L^{\text{Base}} = \prod_{i=1}^{I^{\text{Base}}} P\left[S_i^{\text{Base}} = s_i^{\text{Base}}\right] = \prod_{i=1}^{I^{\text{Base}}} \int_0^1 t^{s_i^{\text{Base}}} (1-t)^{1-s_i^{\text{Base}}} f_R^{\text{Tot}}(t)\,\mathrm{d}t.$$

An equivalent form that lends itself better to numerical optimization is found by using response pattern frequencies $U_r = \{\#_i | S_i = r\}$ to summarize the data. For example, U_3 is the number of parts that is rejected 3 out of K times. For the baseline data, U_1^{Base} is the number of rejects, and $U_0^{\text{Base}} = I^{\text{Base}} - U_1^{\text{Base}}$ is the number of accepted parts. Assuming, as explained before, that the baseline data are collected as a by-product of collecting the sample of rejected parts, we have $I = u_1^{\text{Base}}$. This leads to the following convenient expression for the loglikelihood:

$$\text{logL}(\boldsymbol{\theta}|\boldsymbol{u}) = \sum_{r=0}^{K} u_r \log \int_0^1 f_R^{\text{Tot}}(t|\boldsymbol{\theta})\, t^{r+1} (1-t)^{K-r}\,\mathrm{d}t +$$

$$u_0^{\text{Base}} \log \int_0^1 f_R^{\text{Tot}}(t|\boldsymbol{\theta})\,(1-t)\,\mathrm{d}t, \quad (5.1)$$

where $\boldsymbol{\theta}$ are the parameters of the distribution f_R^{Tot}. We can express *IAP* and *IRP* as functions of the parameters $\boldsymbol{\theta}$ to be estimated:

$$IAP(\boldsymbol{\theta}) = P\left[Y_{ik} = 1 | \tilde{X}_i = 0\right] = P\left[Y_{ik} = 1 | R_i > \frac{1}{2}\right] =$$

$$\frac{\int_{1/2}^{1} (1-r) f_R^{\text{Tot}}(r|\boldsymbol{\theta})\,\mathrm{d}r}{\int_{1/2}^{1} f_R^{\text{Tot}}(r|\boldsymbol{\theta})\,\mathrm{d}r}, \quad (5.2)$$

$$IRP(\boldsymbol{\theta}) = P\left[Y_{ik} = 0 | \tilde{X}_i = 1\right] = P\left[Y_{ik} = 0 | R_i < \frac{1}{2}\right] = \frac{\int_0^{1/2} r f_R^{\mathrm{Tot}}(r|\boldsymbol{\theta})\,\mathrm{d}r}{\int_0^{1/2} f_R^{\mathrm{Tot}}(r|\boldsymbol{\theta})\,\mathrm{d}r}.$$

After maximizing the loglikelihood in (5.1) with respect to $\boldsymbol{\theta}$, (5.2) gives estimates $\widehat{IAP}$ and $\widehat{IRP}$. Further, the (co-)variance matrix can be obtained by applying the delta method on the negative inverse Hessian of the loglikelihood:

$$\begin{aligned}\mathrm{Var}(IAP, IRP) = &\left(\left.\frac{\partial\left(IAP(\boldsymbol{\theta}), IRP(\boldsymbol{\theta})\right)}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right)^T \\ &\times\left(-\left.\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathrm{logL}(\boldsymbol{\theta}|\boldsymbol{u})\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right)^{-1} \times\left(\left.\frac{\partial\left(IAP(\boldsymbol{\theta}), IRP(\boldsymbol{\theta})\right)}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right). \quad (5.3)\end{aligned}$$

5.3 Estimation of f_R^{Tot} by adaptive polynomials

The model in the previous section was specified up to a choice for f_R^{Tot}, the distribution of rejection probabilities R_i of parts. Before we present our choice, we briefly review two alternatives in recent literature, which we will use later for comparison. As shown in Akkerhuis et al. (2016a), Danila et al. (2012, 2013) implicitly assume the following distribution for their method ("Danila model"):

$$\begin{aligned}f_R^{\mathrm{Tot}}(r|\boldsymbol{\theta}) = (1-p) \times \frac{r^{\mu_1\frac{1-\phi_1}{\phi_1}-1}(1-r)^{(1-\mu_1)\frac{1-\phi_1}{\phi_1}-1}}{\mathrm{Beta}\left(\mu_1\frac{1-\phi_1}{\phi_1}, (1-\mu_1)\frac{1-\phi_1}{\phi_1}\right)} + \\ p \times \frac{r^{(1-\mu_0)\frac{1-\phi_0}{\phi_0}-1}(1-r)^{\mu_0\frac{1-\phi_0}{\phi_0}-1}}{\mathrm{Beta}\left((1-\mu_0)\frac{1-\phi_0}{\phi_0}, \mu_0\frac{1-\phi_0}{\phi_0}\right)}.\end{aligned}$$

It is a mixture of two beta distributions, weighted by p, and reparameterized such that the μs are the expected values of the component distributions. The original interpretation is that $p = P[X = 0]$ is the defect rate, and the two component distributions govern the rejection probabilities for conforming and nonconforming parts. Then, μ_0 is the *FAP*, and μ_1 is the *FRP*. The ϕ_0 and ϕ_1 are related to the variance of the beta distributions. The parameters are restricted so that $(1-\mu_0)(1-\phi_0)/\phi_0 > 1$ and $\mu_1(1-\phi_1)/\phi_1 > 1$, because without these restrictions, both components are U-shaped and identification problems occur. In the gold standard unavailable situation, where it is unobservable whether parts are conforming, μ_0 and μ_1 turn out to be poor estimators for *FAP* and *FRP* in general (see Chapters 3 and 4). However, the mixture

distribution f_R^{Tot} could still be fitted to the observed rejection counts S_i, and thus allow estimation of *IAP* and *IRP* via (5.2).

An alternative is given in De Mast et al. (2011) and Erdmann et al. (2016a), and is adopted from item reponse theory (Lord, 1980). It implies that the distribution of R_i is, with Φ the cumulative standard normal distribution,

$$F_R^{\text{Tot}}(r|\boldsymbol{\theta}) = \Phi\left(\frac{1}{\alpha}\log\left(\frac{r}{1-r}\right) + \delta\right),$$

("De Mast model"), a logitnormal distribution (see Chapter 2). It has three extrema: modes close to 0 and 1, and a minimum in between (Frederic and Lad, 2008). The rationale behind this functional form is that parts can be characterized by a standard normally distributed, unobservable property Z. The rejection probability of a part is determined by this property through a logistic characteristic curve: $R_i = q(Z_i)$ with $q(z) = (1 + \exp\{-\alpha(z-\delta)\})^{-1}$. This curve is S-shaped and increases monotonically from 0 to 1. This gives $F_R(r|\boldsymbol{\theta}) = \Phi\left(q^{-1}(r)\right)$. In this model, α is the steepness of the curve and reflects the random error of the measurement system. The logistic curve's inflection point δ is the Z_i-value at which $R_i = \frac{1}{2}$. Again, *IAP* and *IRP* are found from the fitted model using (5.2).

The Danila model has five parameters ($\mu_0, \mu_1, \phi_0, \phi_1$ and p), whereas the De Mast model has only two (α and δ). The latter is more parsimonious, but also relatively inflexible. It has problems fitting situations with $IAP < IRP$ (Erdmann, 2012), and Chapter 4 shows that it easily gives a poor fit when its assumptions are not precisely met. Moreover, it is computationally burdensome as its moments have to be approximated by numerical quadrature. Chapter 4 furthermore shows that the Danila model gives a good fit of rejection probabilities under a wide range of scenarios. Its functional form, however, is strongly tied to its original purpose of estimating *FAP* and *FRP*, and the number of five parameters seems large in view of the fact that only K parameters are needed for a saturated model (see below).

Instead of the options above, we propose to approximate f_R by a product of a squared polynomial (with coefficients $p_0, \ldots, p_\omega$) and the beta distribution:

$$f_R^{\text{Tot}}(r|\boldsymbol{\theta}) = \frac{r^{\alpha-1}(1-r)^{\beta-1}\left(\sum_{i=0}^{\omega} p_i r^i\right)^2}{\sum_{i=0}^{\omega}\sum_{j=0}^{\omega} p_i p_j \text{Beta}(i+j+\alpha, \beta)}, 0 < \alpha, \beta < 1, \boldsymbol{\theta} = (\alpha, \beta, \{p_i\}_{i=0}^{\omega}). \quad (5.4)$$

("Adaptive Polynomial model"). Here, Beta$[.,.]$ is the beta function, the denominator is a normalization constant, and we set $p_0 = 1$ for identifiability. Gallant and Nychka (1987); Stewart (2004) and Bierens (2005) refer to (5.4) and similar models (based

on other distributions than the beta) as "semi-nonparametric". By increasing the polynomial order ω, any pdf on the unit interval can be approximated with arbitrary precision. The number of parameters is increased only when this appears needed for a satisfactory fit.

The beta distribution is very flexible, and allows computation of the integrals in (5.1) without numerical quadrature:

$$\int_0^1 f_R^{\mathrm{Tot}}(r|\boldsymbol{\theta})\, r^{S_i}\,(1-r)^{K-S_i}\,\mathrm{d}r = \frac{\sum_{i=0}^{\omega}\sum_{j=0}^{\omega} p_i p_j \mathrm{Beta}\left(i+j+\alpha+S_i, \beta+K-S_i\right)}{\sum_{i=0}^{\omega}\sum_{j=0}^{\omega} p_i p_j \mathrm{Beta}\left(i+j+\alpha, \beta\right)}.$$

We implement the restriction $0 < \alpha, \beta < 1$, which gives a U-shaped beta distribution. We believe this to be a realistic shape, because probability mass is concentrated at the end points of the unit interval. The interpretation is that most parts are clearly conforming (R_i close to 0) or nonconforming (R_i close to 1), and relatively few parts are more difficult to classify consistently.

There are other options for the base distribution. In preliminary explorations, the Kumaraswamy distribution (Jones, 2009) and uniform distribution were ruled out because the required polynomial order ω for a good fit was higher. Computation of the moments of the bounded Johnson distribution (Johnson, 1949) requires numerical quadrature, leading to long optimization times.

We use the Nelder-Mead algorithm (Nelder and Mead, 1965) to find maximum likelihood estimates. For numerical feasibility, we take $10^{-9} < \alpha, \beta < 1 - 10^{-9}$, because if the maximization algorithm iterates too close to these boundaries, the beta function cannot be calculated. We choose ω adaptively. We start at 0, and only increase if the fit is unsatisfactory. We use a likelihood ratio test on $\boldsymbol{U}$, comparing the observed response pattern frequencies with a saturated model, at a 5% significance level:

$$G = 2\sum_{k=0}^{K} U_k \log \frac{U_k}{E_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}[U_k]} \sim \chi^2\left(K-\omega-2\right).$$

If $G > \chi^2_{0.95}$ then ω is increased by 1 and the likelihood is optimized again. The order ω is not increased beyond $K-3$ to prevent overtfitting, as a perfect fit is possible with $K-2$ degrees of freedom.

5.4 Quantitative evaluation by means of simulation

In this section we evaluate the robustness of the proposed approach and provide sample size recommendations. We use simulation instead of asymptotic theory to

derive properties of the proposed estimators, as the selected polynomial order ω may differ over simulation runs, and this variation should be taken into account when calculating standard errors. In practical situations (in which the true values are unknown, and simulations are thus impossible), asymptotic theory can at least provide approximations to the standard errors.

5.4.1 Robustness

We study the sensitivity of the approach to model misspecification. To this end, we fit the Adaptive Polynomial model while the true data generating mechanisms are instances of the Danila and De Mast models. In each instance, we determine the empirical distribution of the estimators of *IAP* and *IRP* by computing 2500 realizations. For the Danila model, the instances follow a factorial structure with $\mu_0 = 0.02, 0.10, \mu_1 = 0.02, 0.10, \phi_0 = \phi_1 = 0.05, p = 0.05, 0.20$ and $p = 0.125$. The ranges of these values match those in the research by (Danila et al., 2012, 2013). We evaluate the De Mast model in a factorial structure with $\alpha = 5, 12, 20$ and $\delta = 2, 2.5, 3$. The ranges of these values match those of De Mast et al. (2011) and Erdmann et al. (2016a). We use $I = 150$ and $K = 7$ throughout this subsection, which is a typical sample size for binary MSA studies and in line with recommendations such as Danila et al. (2010) and Erdmann et al. (2016a). As explained above, we take the number of baseline data equal to $I^{\text{Base}} = I / [R_i]$.

The results for the Danila model are given in Table 5.1. We find that realizations of the Danila model are approximated well by the Adaptive Polynomial model, with the bias in the estimated *IAP* and *IRP* smaller than one standard error. The largest bias for $\widehat{IAP}$ is 0.0114 (case 3), and for $\widehat{IRP}$ it is 0.0100 (case 4). The modal required polynomial order ω ranges from 2 to 4, corresponding to 4 to 6 model parameters.

The results under the De Mast model are in Table 5.2. The largest bias for $\widehat{IAP}$ is 0.0089, and 0.0004 for $\widehat{IRP}$ (both in case 1). Also in this case, the biases are smaller than the standard errors. The required polynomial order is generally lower.

We conclude that the proposed model is able to fit the data well under a wide range of realistic scenarios. The polynomial order selection works well, and brings expected values of $\widehat{IAP}, \widehat{IRP}$ within one standard error of their true values.

Figure 5.1 (solid curves) shows the density f_R^{Tot} of rejection probabilities implied by the center point cases (row 5 in Tables 5.1 and 5.2). The figure illustrates that typically, most parts either have a rejection probability close to 0 or close to 1. In the Danila model (left graph) there are virtually no parts with rejection probabilities close to 0.5 (that is, the quality of parts seems an almost dichotomous property). The De Mast model implies that the quality of parts is a continuous property, and consequently, there is a fair occurrence of parts that turn out to be difficult to classify consistently (rejection probabilities around $r = 0.5$). The dashed curves demonstrate

	Parameters				True Values		Estimates		$\omega\%$				
	μ_0	ϕ_0, ϕ_1	μ_1	p	IAP	IRP	$\widehat{IAP}$	$\widehat{IRP}$	0	1	2	3	4
1	0.02	0.05	0.02	0.050	0.0200	0.0200	0.0246	0.0195	0	0	44	27	29
							(0.0078)	(0.0029)					
2	0.02	0.05	0.02	0.200	0.0200	0.0200	0.0220	0.0192	1	0	74	6	19
							(0.0055)	(0.0055)					
3	0.02	0.05	0.10	0.050	0.0205	0.1000	0.0319	0.0904	0	0	37	16	46
							(0.0161)	(0.0074)					
4	0.02	0.05	0.10	0.200	0.0201	0.1000	0.0287	0.0900	0	1	11	28	61
							(0.0087)	(0.0104)					
5	0.06	0.05	0.06	0.125	0.0600	0.0600	0.0643	0.0574	0	0	71	4	25
							(0.0115)	(0.0073)					
6	0.10	0.05	0.02	0.050	0.1000	0.0200	0.0988	0.0196	0	0	52	2	47
							(0.0155)	(0.0028)					
7	0.10	0.05	0.02	0.200	0.1000	0.0200	0.1012	0.0193	0	0	64	0	36
							(0.0122)	(0.0054)					
8	0.10	0.05	0.10	0.050	0.1004	0.1000	0.1107	0.0923	0	3	27	20	50
							(0.0266)	(0.0073)					
9	0.10	0.05	0.10	0.200	0.1001	0.1000	0.1033	0.0927	0	0	54	3	43
							(0.0176)	(0.0112)					

Table 5.1: Performance of the adaptive polynomial approach under the Danila model. Values between brackets are the standard errors of the estimates (based on a sample size of $I = 150$ parts and $K = 7$ repeats.

	Parameters		True values		Estimates		$\omega\%$				
	α	δ	IAP	IRP	$\widehat{IAP}$	$\widehat{IRP}$	0	1	2	3	4
1	5	2	0.2154	0.0125	0.2243	0.0121	64	19	8	3	6
					(0.0234)	(0.0017)					
2	5	2.5	0.2370	0.0048	0.2435	0.0047	35	34	17	5	8
					(0.0263)	(0.0006)					
3	5	3	0.2562	0.0015	0.2547	0.0015	6	39	33	10	12
					(0.0297)	(0.0002)					
4	12	2	0.1134	0.0039	0.1169	0.0037	88	4	2	1	5
					(0.0151)	(0.0007)					
5	12	2.5	0.1295	0.0013	0.1337	0.0012	84	6	3	1	5
					(0.0156)	(0.0002)					
6	12	3	0.1447	0.0004	0.1486	0.0003	79	11	3	1	5
					(0.0185)	(0.0001)					
7	20	2	0.0733	0.0022	0.0739	0.0021	88	2	2	1	8
					(0.0114)	(0.0005)					
8	20	2.5	0.0849	0.0007	0.0867	0.0007	88	1	3	1	7
					(0.0126)	(0.0001)					
9	20	3	0.0962	0.0002	0.0984	0.0002	87	2	4	1	6
					(0.0135)	(0.0000)					

Table 5.2: Performance of the adaptive polynomial approach under the De Mast model. Values between brackets are standard errors.

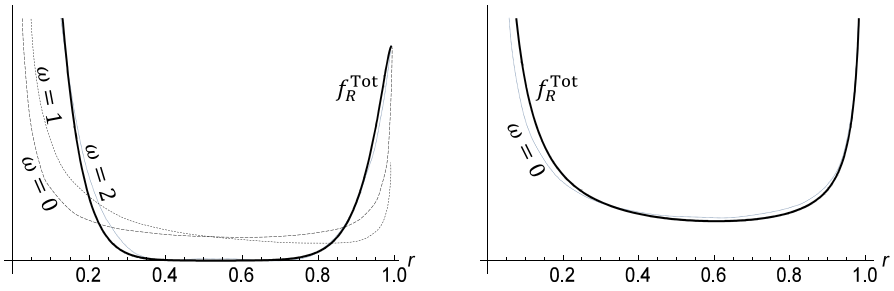


Figure 5.1: Solid line: true pdf of R_i. Dashed lines: adaptive polynomial fits, where smaller dashes indicate higher ω. Left: case 5 for Danila model (required $\omega = 2$); right: case 5 for De Mast model (required $\omega = 0$).

how the densities of rejection probabilities are approximated by the Adaptive Polynomial model, with fits of degrees $\omega = 0, 1, 2$ in the left graph, and $\omega = 0$ in the right graph. Estimates for $1 - IAP$ and IRP are found by taking the means of these fitted curves under the condition that $R_i > 0.5$ and $R_i \leq 0.5$, respectively.

5.4.2 Sample size recommendations

We regard the center point cases (rows 5 in Tables 5.1 and 5.2) as typical situations. With the following parameter values, the proposed Adaptive Polynomial model is nearly identical to these cases: $\alpha = 0.6320, \beta = 0.8868, p_1 = -3.874, p_2 = 3.518$ (to approximate the center point in Table 5.1) and $\alpha = 0.0019, \beta = 0.2796$ (for the center point in Table 5.2). In these points, we have $IAP = 0.0648, IRP = 0.0576$ and $IAP = 0.1368, IRP = 0.0012$ respectively. In order to study the effect of sample size (I and K) on estimator precision, we calculate standard errors of $\widehat{IAP}, \widehat{IRP}$ for $I \in \{50, 60, \ldots, 250\}$ and $K \in \{5, 6, \ldots, 10\}$ in both instances of the Adaptive Polynomial model using 2500 simulation runs per combination of I and K. As before, we take the number of baseline data to be $I^{\text{Base}} = I/E[R_i]$, which it the expected sample size obtained when the baseline data are collected as a by-product of gathering the sample of I rejected parts. When historical baseline data are available, I^{Base} may be larger, and the precision of estimates will be better than the results reported below.

The results are presented in Figure 5.2. To compare various combinations of I and K, dashed lines demarcate a range with a constant total number of measurements $I \times K$. By moving along the dashed line into the region with the lowest standard error, once can find the optimal combination I, K for a given total number of measurements.

The contour lines for the bottom graphs in Figure 5.2 are almost horizontal, i.e.,

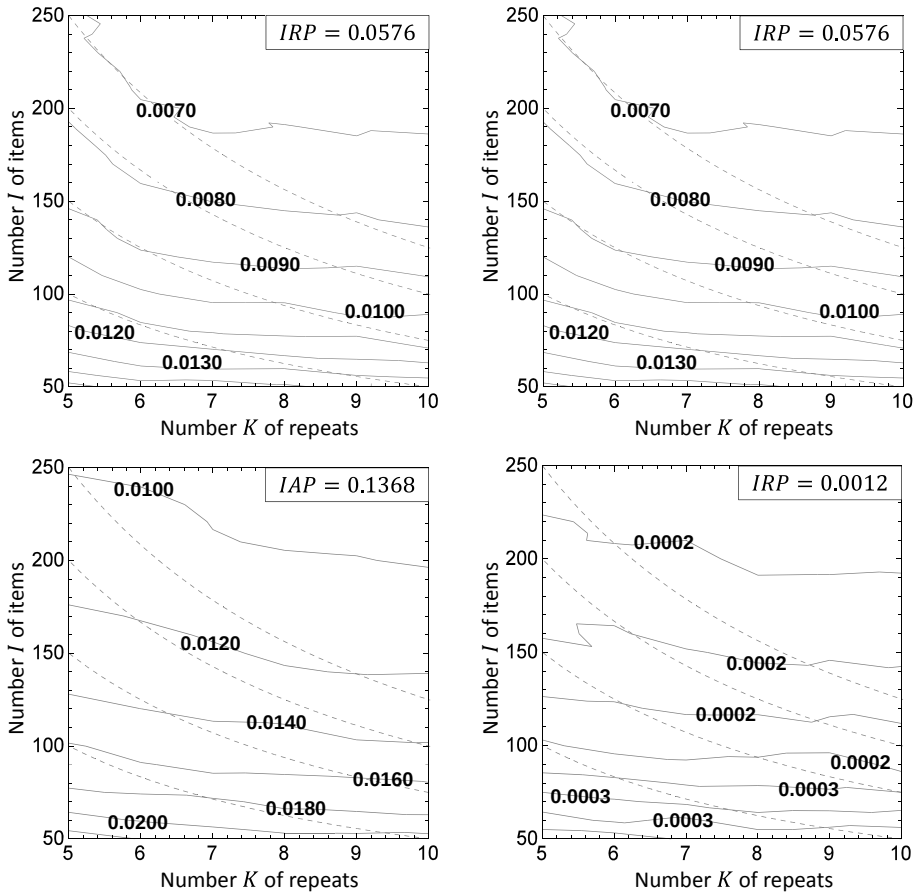


Figure 5.2: Standard errors for $\widehat{IAP}$ (left graphs) and $\widehat{IRP}$ as function of I and K in the points $\alpha = 0.6320, \beta = 0.8868, p_1 = -0.3874, p_2 = 3.518$ (top graphs) and $\alpha = 0.0019, \beta = 0.2796$ (bottom graphs). Dashed curves: region with constant number of measurements $I \times K$.

standard errors are nearly unaffected by K. This means that standard errors for both $\widehat{IAP}$ and $\widehat{IRP}$ are minimized by exchanging repeats for a larger sample. Also in the top graphs the contours indicate that larger I and smaller K is optimal, except that for estimating IAP, the number K of repeats should not get below 6 (as the standard errors increase sharply when going from $K = 6$ to $K = 5$). This can be related to the value of ω (which may go up to 4), which, in combination with $K \leq \omega + 2$, gives a saturated model. We recommend taking $K = 7$ and I in the range 100 to 150, which gives standard errors below 25% of the true IAP and IRP. These sample sizes are augmented with as many useable baseline data that happen to be available (with $I^{\text{Base}} = I/E\left[R_i\right]$ as a minimum).

5.5 Example: optical inspection of car parts

We have data from a car parts manufacturer (also see Chapter 2). The measurement system under consideration is an automated optical inspection that judges the soldering of a sensor component as "OK" or "NOK". The size of the experiment was set to $I \times K = 150 \times 7$ (see columns headed OBSERVED in Table 5.3). From the 150 parts in the sample, 110 were rejected in all 7 repeats, and none of the parts was rejected fewer than 2 times (note that this does not reflect the quality of the production process, since this sample was drawn from the stream of rejected parts). There were many more baseline data than the measurements needed to obtain 150 rejects, as the inspection system had been in use for some time in this high-volume production process. We were assured by the manufacturer that nothing had changed to make these baseline data unrepresentative for the current situation. The historical reject rate was $1,271/254,200 = 0.0050$.

Fitting f_R^{Tot} to the data with $\omega = 0$ gave the results in the first row of Table 5.4. The G-test rejected the fit. ω was increased to 1, the fit was not rejected, and the result is given in the second row of Table 5.4 (standard errors computed using (5.3)). Fitted response pattern frequencies are given in Table 5.3 (see columns headed FITTED).

For comparison, the Danila and De Mast models were fitted as well. This gave problems. The estimation algorithm did not converge for the Danila model. Also gradient-based optimization did not converge, and neither did the "ad hoc procedure" described in Danila et al. (2012), which is to decrease the number of parameters by setting $\phi_0 = \phi_1 = \phi$. We found that convergence failure was caused by $\widehat{\mu}_1$ going to zero. This might be caused by the absence of parts with 0 or 1 rejections. We coped with the problem by posing the restriction $\mu_1 > 10^{-10}$. The resulting fit yields the estimates in Table 5.4 (third row), which are close to those obtained from the Adaptive Polynomial model (order 1). The estimation algorithm based on the De Mast model did converge, but the fit is significantly rejected.

	OBSERVED		FITTED	
# rejects	Sample	Baseline data	Sample	Baseline data
0	0	252,929	2.46	252,929
1	0	1,271	2.82	1,271
2	1		3.30	
3	6		4.01	
4	6		5.15	
5	6		7.30	
6	21		13.1	
7	110		112	
Totals	150	254,200		

Table 5.3: Car parts example: observed and fitted response pattern frequencies u_r ($r = 0, 1, \ldots, 7$) and u_r^{Base} ($r = 0, 1$).

Model	$\widehat{IAP}$	$\widehat{IRP}$	G-test
Adap. Pol.	0.0694	0.0004	$G = 18.081, df = 6,$
($\omega = 0$)	(0.0096)	(0.0001)	$p = 0.0060$
Adap. Pol.	0.0778	0.0001	$G = 6.3123, df = 5,$
($\omega = 1$)	(0.0107)	(0.0000)	$p = 0.2770$
Danila	0.0778	0.0001	$G = 6.2387, df = 3,$
	(0.0107)	(0.0001)	$p = 0.1006$
De Mast	0.0664	0.0004	$G = 19.986, df = 6,$
	(0.0090)	(0.0001)	$p = 0.0028$

Table 5.4: Four models fitted to the car parts data, and resulting estimates for *IAP* and *IRP* (standard errors between brackts).

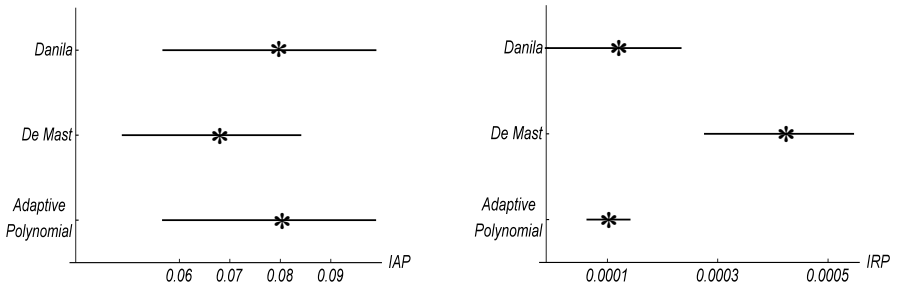


Figure 5.3: Point estimates and 95% confidence intervals for various estimators of *IAP* and *IRP*.

Figure 5.3 provides a graphical overview of the estimates and their 95% confidence intervals. In general, we can conclude that the proposed method gives relatively good confidence interval widths. Moreover, the location of the confidence intervals is plausible, as they fully overlap with the confidence intervals of the most flexible model, namely, the Danila model. Note that the fitted Adaptive Polynomial model is more parsimonious than the Danila model (3 versus 5 parameters).

From the fitted Danila model we can also derive estimates for FAP and FRP, as this was its original purpose in Danila et al. (2012, 2013). These are $\widehat{FAP} = 0.1108$ and $\widehat{FRP} = 0.0000$. We find it hard to see how these estimates are meaningful values. Theoretically, the estimated FAP (say) is the fraction of nonconforming parts that is falsely accepted, but with conformance being an unobservable and only vaguely defined concept without numerical specifications, we are inclined to believe that the concept of a false acceptance, and hence the produced estimate for FAP, are difficult to relate back to a relevant property of the inspection system. This reflects the general point made in this chapter, that estimation of FAP and FRP is problematic without gold standard.

We argue that the determination of a test's IAP and IRP is a useful alternative. If IAP and IRP are substantial, this implies that the test cannot consistently discriminate worse from better parts, and this disqualifies the test system for an application. If IAP and IRP are reasonably good, the system's discriminative capability is sufficient for the application. If possible, it would be desirable in that case to rule out that the test results are systematically too strict or too lenient. Such would require a calibration of the system. This remains problematic without a gold standard, and a rough calibration based on consensus among engineers may be the best one can do.

5.6 Discussion and conclusion

The false acceptance probability (FAP) and false rejection probability (FRP) of binary tests cannot be estimated reliably if a gold standard is not available, as shown convincingly by mathematical analysis in Chapter 3 and by simulation in Chapter 4. In that case, only the random component of measurement error is estimable, as quantified by the inconsistent acceptance probability (IAP) and inconsistent rejection probability (IRP). In the absence of systematic measurement error, these probabilities equal FAP and FRP.

In this chapter we outlined a method for estimating IAP and IRP. Data are gathered by collecting I rejected parts, and evaluating them K times, and augmented with baseline data. A probability distribution of rejection probabilities, f_R^{Tot}, is fitted to these data, which is used to calculate $\widehat{IAP}$ and $\widehat{IRP}$. We propose a functional form of f_R^{Tot} that is adaptive, meaning that its flexibility can be increased in steps as long

as a goodness-of-fit test rejects the fit. In theory, the model should be able to fit any situation. Indeed, our simulation studies show that biases are small.

An application in a real case exemplifies the unreliability of current approaches, and the usefulness of the proposed method in practice. The case study illustrates that often, estimation of *FAP* and *FRP* is unattainable without gold standard, and that the relative inflexibility of the approach of Erdmann et al. (2016a) (presented in Chapter 2) limits its applicability for estimating *IAP* and *IRP*. The approach presented in this paper seems the most promising candidate for solving this problem so far.

Chapter 6

Probabilities of inconsistent classification for nominal measurements

Chapters 3 and 4 have analyzed the problem of measurement system analysis for binary data when a gold standard is not available, and this culminated in the method presented in Chapter 5. We believe that, thus, we have fulfilled the first two goals of the thesis.

The third goal of this thesis is to generalize its findings regarding binary measurements to the case of measurement on a scale with more than two classes. Following the reasoning in Chapter 3, we split measurement error in a systematic and a random component. As in Chapter 5, a method is offered to estimate the random component. As a generalization of the inconsistent acceptance and rejection probabilities (IAP and IRP) from the previous chapters, this chapter introduces the more general Inconsistent Classification Probability (ICP). This chapter ends with a case study in the last section.

6.1 Introduction

Nominal measurement systems are devices, persons, or persons operating devices, that classify objects on a nominal scale. They yield a measurement outcome Y, which takes on one of C possible values (categories or classes). An example from industry (the context of this thesis) is an operator who assigns a failure mode to products that failed quality inspection. Since many decisions are a result of measurement,

a measurement system should be accompanied by a quantification of its reliability in terms of measurement error. The experiment and subsequent analysis needed to arrive at such quantification is called "measurement system analysis" (MSA). In this chapter, we look at MSA for nominal measurement systems, or "nominal MSA".

In symbols, measurement error is often expressed as the event $\{Y = a|X \neq a\}$, with Y a measurement outcome and X the true value. X is obtained by applying a "reference standard" or "gold standard": an authoritative procedure to arrive at the true value of the object. We consider the case that this standard is not available: the "gold standard unavailable case". This may be because a gold standard is too costly to apply, has a negative impact on the object under measurement, or simply does not exist, for example for lack of a clear definition.

Past literature about nominal MSA has focused on the concept of agreement, which originated in the social sciences (Cohen, 1960). Agreement is the event that two measurements on the same object i yield equal outcomes, $\{Y_{i1} = Y_{i2}\}$. A popular measure of agreement is the kappa (κ) statistic (Fleiss, 1971) which can be found in industry standards such as AIAG (2003) and medical literature (Erdmann et al., 2016b). Log-linear modeling is a related approach (Tanner and Young, 1985; Agresti, 1988). There have been many criticisms regarding agreement studies (De Mast, 2007; Erdmann et al., 2016a), and moreover, agreement quantifies a different concept than measurement error: even if appraisers agree on some decision, it is not necessarily the correct decision X.

Proneness to measurement error can be quantified using probabilities of false classification, $FCP_a = P[Y = a|X \neq a]$. These are the probabilities that measurement of an object gives a while the true value of this object is not a. For the binary case, Akkerhuis et al. (2016a,b) have shown that false classification probabilities are in general not identifiable without gold standard, even though many attempts have been made. The goal of this chapter is to quantify the random component of the FCPs in terms of probabilities of "inconsistent classification" (ICPs) on scales with more than two classes.

We define the probabilities of inconsistent classification as the conditional probabilities $ICP_a = P\left[Y = a|\tilde{X} \neq a\right]$. These condition on the most likely outcome $\tilde{X}_{ij}$ of a measurement of object i by appraiser j. That is, $\tilde{X}_{ij}$ is the measurement outcome that object i would receive most frequently from appraiser j during repeated measurement. This is comparable to common practice for numerical MSA, where random measurement error is conceived as deviation from the expected measurement outcome as opposed to deviation from the true value. In the nominal case, random variables do not have a statistical expectation, and we replace it by the mode. Since it does not condition on a true value, but on something that is observable, it is an identifiable property.

There are other ways to quantify random measurement error. In some contexts,

a quantification like $P\left[Y \neq a|\tilde{X} = a\right]$ or even $P\left[\tilde{X} = a|Y \neq a\right]$ or $P\left[\tilde{X} = a|Y \neq a\right]$ may be useful as well. The modeling we propose will allow calculation of any of these probabilities, but for conciseness, we focus on $ICP_a = P\left[Y = a|\tilde{X} \neq a\right]$ because it is the most natural extension of the Inconsistent Acceptance/Rejection Probabilities in binary MSA (Chapter 3): probabilities of a specific outcome, conditional on the event that this outcome is Inconsistent.

Akkerhuis et al. (2016a) have developed a method to estimate ICPs in the binary case, and this chapter presents an extension. Estimation of ICPs for ordinal MSA has been explored by De Mast and Van Wieringen (2010), but their approach relies on the assumption of a univariate continuum underlying the measurement scale, which is not reasonable for unordered, nominal scales.

In Section 6.2, we introduce notation, modeling, estimation and diagnostics. We illustrate how ICPs represent random components of FCPs. Section 6.3 explores the statistical properties of the proposed estimators, and Section 6.4 demonstrates the method using data from a real life case.

6.2 Statistical modeling

In the first subsection, notation and modeling of the measurement outcomes are introduced. The second subsection covers estimation of parameters, and consequently, of the ICPs, followed by diagnostic tests in the third subsection. The final subsection illustrates what it means that ICPs are random components of the FCPs.

6.2.1 Modeling of measurement outcomes

In a typical MSA experiment, I objects are measured K times by J appraisers on a scale with C classes, leading to outcomes $\boldsymbol{Y} = \{Y_{ijk}\}_{i=1,\ldots,I;j=1,\ldots,J,k=1,\ldots,K}$.

Objects are modeled by probability vectors $\boldsymbol{P}_{ij}$ that contain classification probabilities. Conditional on a realization of this vector $\boldsymbol{p}_{ij}$, the repeated measurements of object i by appraiser j follow a multinomial distribution with parameters K and $\boldsymbol{p}_{ij}$. Note that $\boldsymbol{P}_{ij}$ is indexed by j to indicate that the distribution of appraisals of some object i may be different for each of the J appraisers.

In symbols, we have

$$\{Y_{ijk}|\, \boldsymbol{P}_{ij} = \boldsymbol{p}_{ij}\}_{k=1,\ldots,K} \sim MN\left(K, \boldsymbol{p}_{ij}\right),$$
$$\boldsymbol{p}_{ij} = \{p_{ij1}, \ldots, p_{ijC}\}, \sum_{c=1}^{C} p_{ijc} = 1,$$

For instance, $\boldsymbol{p}_{ij} = \{0.01, 0.98, 0.01\}$ means that appraiser j mostly classifies object i as 2, but sometimes inconsistently as 1 or 3. If $\boldsymbol{p}_{ij} = \{0, 1, 0\}$, the appraiser will

classify i as 2 with certainty. If $\boldsymbol{p}_{ij} = \{0.01, 0.49, 0.50\}$, there is doubt between the categories 2 and 3, but classification 1 is almost never given. The modal outcome $\tilde{X}$ is defined as the position of the largest element, that is, $\tilde{X}_{ij} = \arg\max_c p_{ijc}$.

We model the $\boldsymbol{P}_{ij}$ as draws from a probability distribution $F_{j,\boldsymbol{P}}(p_{ij1}, \ldots, p_{ijC}) = P[P_{ij1} \leq p_{ij1}, \ldots, P_{ijC} \leq p_{ijC}]$. The Dirichlet distribution is a natural choice, as it is often used as a prior for multinomial data in Bayesian statistics and allows for explicit calculations. Its density function is given by:

$$f_{j,\boldsymbol{P}}(p_{ij1}, \ldots, p_{ijC}) = \frac{\Gamma\left(\sum_{c=1}^{C} \beta_j \alpha_{jc}\right)}{\prod_{c=1}^{C} \Gamma(\beta_j \alpha_{jc})} \prod_{c=1}^{C} p_{ijc}^{\beta_j \alpha_{jc} - 1} \text{ for } \sum_{c=1}^{C} p_{ijc} = 1$$
$$\alpha_{jc} > 0, \sum_{c=1}^{C} \alpha_{jc} = 1, \beta_j > 0.$$

This distribution $F_{j,\boldsymbol{P}}$ can be conceived as a mixture distribution $\sum_{c=1}^{C} P[X = c] \times F^c_{j,\boldsymbol{P}}$, where the components $F^c_{j,\boldsymbol{P}} = P[P_{ij1} \leq p_{ij1}, \ldots, P_{ijC} \leq p_{ijC} | X = c]$ are the distributions of $\boldsymbol{P}_{ij}$ amongst objects with true value c. However, as both X and its distribution are unknown, only the aggregate distribution is identifiable, and not its components.

Where the $\boldsymbol{P}_{ij}$ are the expected relative frequencies of classifications of an object i (that is, $p_{ijc} = P\left[Y_{ijk} = c | \boldsymbol{P}_{ij} = \boldsymbol{p}_{ij}\right]$), the $\boldsymbol{\alpha}_j$ are the expected relative frequencies in the population of objects: $\alpha_{jc} = E[P_{ijc}] = E_{\boldsymbol{P}_{ij}}\left[P\left[Y_{ijk} = c | \boldsymbol{P}_{ij} = \boldsymbol{p}_{ij}\right]\right]$.

The scalar β_j is a concentration parameter. For $\beta_j \to 0$, probability mass in $f_{j,\boldsymbol{P}}$ will concentrate only around the C unit vectors $(1, 0, 0, \ldots)$, $(0, 1, 0, \ldots)$, etc. The realizations of $\boldsymbol{P}_{ij}$ will then almost surely all be unit vectors. Consequently, the distribution of $\left\{Y_{ij1}, \ldots, Y_{ijK} | \boldsymbol{P}_{ij} = \boldsymbol{p}_{ij}\right\}$ becomes degenerate, leading to perfectly consistent classifications. The C unit vectors are drawn in proportions given by $\boldsymbol{\alpha}_j$.

If, however, $\beta_j \to \infty$, probability mass in $f_{j,\boldsymbol{P}}$ will concentrate only around $\boldsymbol{\alpha}_j$, leading to realizations of $\boldsymbol{P}_{ij}$ that are almost identical (and equal to $\boldsymbol{\alpha}_j$) for each object i. This means that measurement outcomes do not at all depend on the objects under study, and we call these measurements "uninformative". In conclusion, a low β_j implies better consistency of measurements, while a high β_j means poor consistency and uninformative measurements.

If $\beta_j \to \infty$, and measurements are uninformative, categorization becomes purely random guessing. Naturally, we would expect the distributions $\boldsymbol{P}_{ij}$ then to approach uniform distributions. That is, if $\beta_j \to \infty$, we would expect to find that approximately $\boldsymbol{\alpha}_j = \left(\frac{1}{C}, \ldots, \frac{1}{C}\right)$. This is the maximum entropy and thus maximally uninformative distribution for the Y_{ijk} (De Mast, 2007). However, due to some artifact in the measurements or due to behavioral tendencies of the appraisers, it may be that purely

random categorizations have a non-uniform distribution (that is, $\beta_j \to \infty$ while $\boldsymbol{\alpha}_j \neq \left(\frac{1}{C}, \ldots, \frac{1}{C}\right)$).

6.2.2 Estimation of ICPs

We propose estimation of the parameters in $f_{j,\boldsymbol{P}}$ by maximum likelihood. Conditional on a realization $\boldsymbol{p}_{ij}$, the outcome of repeated measurement of object i by appraiser j has likelihood

$$P\left[Y_{ij1} = y_{ij1}, \ldots, Y_{ijK} = y_{ijK} | \boldsymbol{P}_{ij} = \boldsymbol{p}_{ij}\right] = \prod_{c=1}^{C} p_{ijc}^{\#_k\{Y_{ijk}=c\}}.$$

Here $\#_k\{Y_{ijk} = c\}$ is the number of occurrences of classification c in repeated measurements $\{Y_{ij1}, \ldots, Y_{ijK}\}$, and we assume that Y_{ijk} are independent conditional on a realization $\boldsymbol{p}_{ij}$. In symbols:

$$\left\{Y_{ijk} | \boldsymbol{P}_{ij} = \boldsymbol{p}_{ij}\right\}_{i=1,\ldots,I, j=1,\ldots,J, k=1,\ldots,K} \text{ are independent.}$$

This assumption has the interpretation that, besides the $\boldsymbol{P}_{ij}$, there are no factors that induce dependencies amongst the $Y_{ijk}, k = 1, \ldots, K$. As $\boldsymbol{P}_{ij}$ is not observed directly, we integrate it out over its probability distribution, and obtain the unconditional likelihood:

$$P\left[Y_{ij1} = y_{ij1}, \ldots, Y_{ijK} = y_{ijK}\right] = \int_0^1 \ldots \int_0^1 f_{j,\boldsymbol{P}}\left(p_{ij1}, \ldots, p_{ijC}\right) \times$$

$$\prod_{c=1}^{C} p_{ijc}^{\#_k\{Y_{ijk}=c\}} \mathrm{d}p_{ij1} \ldots \mathrm{d}p_{ijC} = \frac{\Gamma\left(\sum_{c=1}^{C} \beta_j \alpha_{jc}\right)}{\prod_{c=1}^{C} \Gamma\left(\beta_j \alpha_{jc}\right)} \times \frac{\prod_{c=1}^{C} \Gamma\left(\beta_j \alpha_{jc} + \#\left\{Y_{ijk} = c\right\}\right)}{\Gamma\left(\sum_{c=1}^{C} \beta_j \alpha_{jc} + \#_k\left\{Y_{ijk} = c\right\}\right)}.$$

The loglikelihood is obtained by aggregating log-probabilities over objects:

$$L_j\left(\left\{\alpha_{jc}\right\}_{c=1,\ldots,C}, \beta_j\right) = \sum_{i=1}^{I} \log \frac{\Gamma\left(\sum_{c=1}^{C} \beta_j \alpha_{jc}\right)}{\prod_{c=1}^{C} \Gamma\left(\beta_j \alpha_{jc}\right)} \times \frac{\prod_{c=1}^{C} \Gamma\left(\beta_j \alpha_{jc} + \#\left\{Y_{ijk} = c\right\}\right)}{\Gamma\left(\sum_{c=1}^{C} \beta_j \alpha_{jc} + \#_k\left\{Y_{ijk} = c\right\}\right)}. \tag{6.1}$$

The number of parameters is $C + 1$ (one of which is trivial as the C elements in $\boldsymbol{\alpha}$ sum to 1) for each appraiser. We found in our analyses that the Nelder-Mead algorithm (Nelder and Mead, 1965) gives stable estimates. Instead of maximizing L_j

with respect to $\alpha_{j1}, \ldots, \alpha_{jC}, \beta_j$ under the restriction that $\sum_c \alpha_{jc} = 1$, we recommend to maximize L_j with respect to $\alpha_{j1}\beta_j, \ldots, \alpha_{jC}\beta_j$.

ICP_{ja} can be calculated using the indicator function (denoted $\mathbf{1}$):

$$ICP_{ja} = P\left[Y_{ijk} = a | \tilde{X}_{ij} \neq a\right] = \frac{P\left[Y_{ijk} = a\right] - P\left[Y_{ijk} = a, \arg\max_c p_{ijc} = a\right]}{1 - P\left[\arg\max_c p_{ijc} = a\right]} = \frac{\int\limits_S f_{j,\boldsymbol{P}}\left(p_{ij1}, \ldots, p_{ijC}\right) \mathbf{1}_{\{\arg\max_c p_{ijc} = a\}} \mathrm{d}\boldsymbol{p}_{ij}}{1 - \int\limits_S f_{j,\boldsymbol{P}}\left(p_{ij1}, \ldots, p_{ijC}\right) \mathbf{1}_{\{\arg\max_c p_{ijc} = a\}} \mathrm{d}\boldsymbol{p}_{ij}} - \frac{\int\limits_S f_{j,\boldsymbol{P}}\left(p_{ij1}, \ldots, p_{ijC}\right) p_{ija} \mathbf{1}_{\{\arg\max_c p_{ijc} = a\}} \mathrm{d}\boldsymbol{p}_{ij}}{1 - \int\limits_S f_{j,\boldsymbol{P}}\left(p_{ij1}, \ldots, p_{ijC}\right) \mathbf{1}_{\{\arg\max_c p_{ijc} = a\}} \mathrm{d}\boldsymbol{p}_{ij}}.$$

These C integrals are not easily approximated by numerical quadrature due to the complex intersection of the support S of $f_{j,\boldsymbol{P}}$ and the range of $\boldsymbol{P}_{ij}$ that satisfies the indicator condition. We therefore recommend the following transformation and corresponding Jacobian determinant:

$$p_{ijk} = \begin{cases} \frac{p^*_{ijk}}{1 + \sum_{c=1}^{C} p^*_{ijc} - p^*_{ija}} & k \neq a \\ \frac{1}{1 + \sum_{c=1}^{C} p^*_{ijc} - p^*_{ija}} & k = a \end{cases}, \det\left[\frac{\partial p_{ijk}}{\partial p^*_{ijk}}\right] = \frac{1}{\left(1 + \sum_{c=1}^{C} p^*_{ijc} - p^*_{ija}\right)^2}$$

By integrating $\left\{p^*_{ijc}\right\}_{c=1}^{C}$ (with the exception of p^*_{ija}) over the unit hypercube, conditions $\arg\max_c p_{ijc} \neq a$ and $\sum_{c=1}^{C} p_{ijc} = 1$ are automatically satisfied.

For $C > 5$, standard adaptive numerical quadrature turned out computationally prohibitive, and Monte Carlo integration provided a better alternative.

6.2.3 Diagnostics and tests

To verify the fit of the model, response patterns $\boldsymbol{U}_{ij} = \left(\#_k\left\{Y_{ijk} = c\right\}\right)_{c=1,\ldots,C}$ are defined. An example of such a pattern is $\boldsymbol{u}_{ij} = (1, 0, 2)$: object i is classified once as 1 and twice as 3 in three repeated measurements by appraiser j.

The data from the MSA study are summarized in response pattern frequencies E:

$$E_j\left(\boldsymbol{u}\right) = \#_i\left\{\boldsymbol{U}_{ij} = \boldsymbol{u}\right\}.$$

For example, $E_j\left((0, 2, 0)\right) = 3$ means that 3 objects are classified as 2 consistently by appraiser j. Goodness-of-fit is verified by comparing E_j to the expected frequencies

(in which $u[c]$ is the cth element of $\boldsymbol{u}$):

$$E\left[E_j\left(\boldsymbol{u}\right)\right] = I \times \int_0^1 \dots \int_0^1 e_j\left(p_{ij1}, \dots, p_{ijC}\right) \frac{K!}{\prod_{c=1}^{C} u[c]!} \times \prod_{c=1}^{C} p_{ijc}^{u[c]} \mathrm{d}p_{ij1} \dots \mathrm{d}p_{ijC} =$$

$$I \times \frac{K!}{\prod_{c=1}^{C} u[c]!} \frac{\Gamma\left(\sum_{c=1}^{C} \beta_j \alpha_{jc}\right)}{\prod_{c=1}^{C} \Gamma\left(\beta_j \alpha_{jc}\right)} \frac{\prod_{c=1}^{C} \Gamma\left(\beta_j \alpha_{jc} + u[c]\right)}{\Gamma\left(\sum_{c=1}^{C} \beta_j \alpha_{jc} + u[c]\right)}.$$

Note that in a high variety of possible response patterns, expected frequencies are often low, which makes the standard chi-square test unreliable as shown by Kallenberg et al. (1985). We therefore use the likelihood ratio test, which is also known as the G-test when it comes to fitting response pattern frequencies:

$$G_j = 2 \sum_{\text{all}\boldsymbol{u}} E_j\left(\boldsymbol{u}\right) \times \log \frac{E_j\left(\boldsymbol{u}\right)}{E\left[E_j\left(\boldsymbol{u}\right)\right]} \sim \chi^2\left(\#\text{ unique } \boldsymbol{U} - 1 - C\right).$$

$\#$ unique $\boldsymbol{U} - 1$ is the number of nontrivial parameters needed for a saturated model, and C is the number of parameters in the proposed model (although there are $C + 1$ parameters, one is trivial because of the restriction $\sum \alpha_c = 1$). The estimates for $\boldsymbol{\alpha}_j$ and β_j can be used to test against uninformative categorization, or to compare appraisers.

- To test for uninformative categorization, one can test $H_0 : f_{j,\boldsymbol{P}} = f_{j,\boldsymbol{P}}^{\infty}$ with $f_{j,\boldsymbol{P}}^{\infty} = \lim_{\beta_j \to \infty} f_{j,\boldsymbol{P}}$. To, additionally, test for random categorization, one can include $\boldsymbol{\alpha}_j = \left\{\frac{1}{C}\right\}_{c=1}^{C}$ in the null hypothesis.
- A comparison of appraisers j and k with respect to consistency can be performed by testing $H_0 : \beta_j = \beta_k$ for $j \neq k$.
- Systematic differences between appraisers can be detected by testing $H_0 : \boldsymbol{\alpha}_j = \boldsymbol{\alpha}_k$ for $j \neq k$.
- Interchangeability of appraisers can be tested by $H_0 : \boldsymbol{\alpha}_j = \boldsymbol{\alpha}_k, \beta_j = \beta_k$ for $j \neq k$.

The likelihood ratio test can be used for these purposes, where the likelihood (6.1) is optimized both with and without these restrictions. However, no appraiser is the same, and thus, for large enough samples, we will always find significant differences, however small. Because statistical significance does not imply practical relevance, we

recommend to combine hypothesis testing with a judicious look at the actual sizes of the differences.

For evaluating the actual sizes of these differences, it can help to construct and plot confidence intervals. By calculating minus the inverse of the Hessian matrix of the loglikelihood and then taking the square roots of the diagonal elements, we find standard errors. 95% confidence intervals are roughly given by the estimated value ± 1.96 times the standard error.

6.2.4 Probabilities of mis- and inconsistent classification

As claimed in the introduction, probabilities of inconsistent classification ICP_{ja} are the random components of the probabilities of false classification FCP_{ja}. We show the mathematical interpretation by following the reasoning in Akkerhuis et al. (2016a).

The modal outcome $\tilde{X}_{ij}$ allows decomposing the probability of misclassification FCP_{ja} as follows:

$$FCP_{ja} = P\left[Y_{ijk} = a, \tilde{X}_{ij} = a | X_i \neq a\right] + P\left[Y_{ijk} = a, \tilde{X}_{ij} \neq a | X_i \neq a\right].$$

The first term, $P\left[Y_{ijk} = a, \tilde{X}_{ij} = a | X_i \neq a\right]$, is the systematic component, where the modal outcome does not equal the true value, but the actual outcome does equal the modal outcome. In $P\left[Y_{ijk} = a, \tilde{X}_{ij} \neq a | X_i \neq a\right]$ (the second term), we have that the measurement outcome randomly deviates from the modal outcome. The second term can be further decomposed as follows:

$$P\left[Y_{ijk} = a, \tilde{X}_{ij} \neq a | X_i \neq a\right] = P\left[Y_{ijk} = a | \tilde{X}_{ij} \neq a\right] \times c_1 - P\left[Y_{ijk} = a, \tilde{X}_{ij} \neq a | X_i = a\right] \times c_2.$$

With $c_1 = P[\tilde{X} \neq a]/P[X \neq a]$ and $c_2 = P[X = a]/P[X \neq a]$. This decomposition reveals an interesting fact that does not hold for numerical measurements. The first part is merely random error. However, a probability is subtracted for the event that a measurement outcome Y_{ijk} equals the true value X_i, but $\tilde{X}_{ij} \neq X_i$. That is, a random error cancels out a systematic error, an event that has zero probability for numerical measurements.

Consequently, we use the probability of inconsistent classification ICP_{ja} to quantify random measurement error: $ICP_{ja} = P\left[Y_{ijk} = a | \tilde{X}_{ij} \neq a\right]$. Note that in the absence of systematic measurement error ($\tilde{X}_{ij} = X_i$), we have that $ICP_{ja} = FCP_{ja}$.

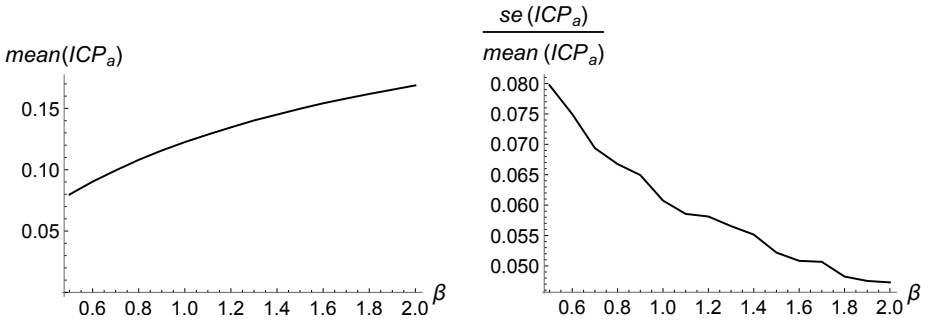


Figure 6.1: Average ICP_a (left), and relative standard error of ICP_a (right) for different values of β.

6.3 Evaluation by means of simulation

In this section, we evaluate first and second order properties of the proposed estimators of ICP_a by means of simulation. We vary the concentration parameter β and the number of classes C. We also investigate the effect of non-uniformity in $\boldsymbol{\alpha}$, to represent cases in which some classes are rare. Finally, we evaluate the relation between sample size I and number of repeated measurements K on the one hand, and standard errors of ICP_a on the other.

For ease of notation, in what follows, we omit subscripts j that index appraisers.

6.3.1 Effect of different concentration β

We study the situation of classifications on a three-point scale. For the case $\boldsymbol{\alpha} = \left\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right\}$ we vary $\beta = 0.50, 0.55, \ldots, 2.00$ (high concentration to low concentration). In particular, for each value of β, we perform 1000 simulation runs in which we draw $I = 300$ realizations $\boldsymbol{p}_i$ from the Dirichlet distribution, perform $K = 10$ repeated measurements for each object i, optimize the likelihood to obtain estimates for $\widehat{\boldsymbol{\alpha}}, \widehat{\beta}$, and use these to calculate $\widehat{ICP}_a$. This gives us the empirical distribution of the estimators for each value of β.

Figure 6.1 contains plots of the average estimates and the associated standard errors, where interpolation between the points has been used. In the left plot, we see that, naturally, the higher β becomes (the lower the concentration), the higher the probability of inconsistent classification. For low values of β, draws for $\boldsymbol{P}_{ij}$ are close to unit vectors, leading to highly consistent classifications. For high values of β, $\boldsymbol{p}_{ij}$ is closer to $\boldsymbol{\alpha}_j$.

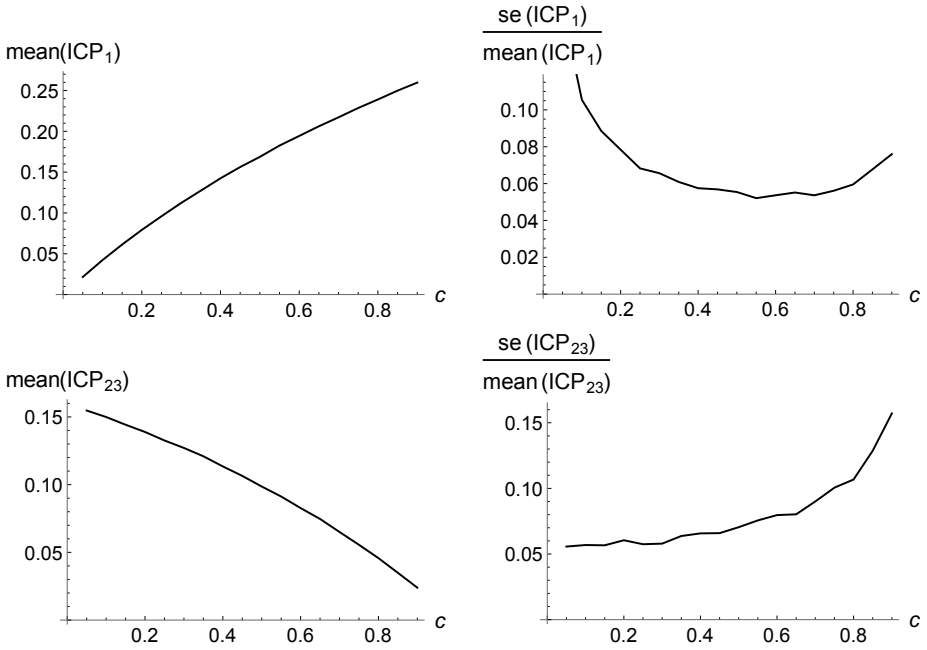


Figure 6.2: Average ICP_a (left), and relative standard error of ICP_a (right) for different values of q. Top row: ICP_1; bottom row: ICP_2 and ICP_3.

The right plot in Figure 6.1 shows that the relative standard error of the estimates is decreasing. Higher values of β can be estimated relatively precisely. This means that the mean increases faster in β than does the standard error.

6.3.2 Effects of non-uniformities in $\boldsymbol{\alpha}$

We consider a variety of situations in which $\boldsymbol{\alpha}$ is not uniform. In particular, we consider $\boldsymbol{\alpha} = \left\{q, \frac{1-q}{2}, \frac{1-q}{2}\right\}$ for $q = 0.05, 0.10, \ldots, 0.90$. For low values of q, there is one underrepresented class, and for high values of q, there are two underrepresented classes.

For each q, we perform 1000 simulation runs to generate the empirical distribution of $ICP_a; a = 1, 2, 3$. We set $I = 300$, $K = 10$ and $\beta = 1$. Figure 6.2 shows, as a function of q, the means of the estimators of ICP_1 and ICP_2 (and ICP_3, which equals ICP_2 up to some simulation error), as well as their relative standard errors.

We see, first, that the average misclassification probability in favor of category 1 increases when q increases (top graphs in Figure 6.2). Put simpler: as α_1 increases, the

$\boldsymbol{\alpha}$	$\left(\frac{1}{2}, \frac{1}{2}\right)$	$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$	$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$
mean ICP_a	0.1823	0.1225	0.0924
rel. s.e. ICP_a	0.0107	0.0076	0.0064

Table 6.1: Average ICP_a (left), and relative standard error of ICP_a (right) for different values of C.

probability of misclassifying something as 2 or 3 decreases (bottom graphs in Figure 6.2). This makes sense: if an appraiser thinks there are many 1s in the population, he or she will be more inclined to misclassify 2s or 3s as 1, and vice versa.

Second, from the standard errors, we see that the estimate for ICP_1 is most precise if $q \approx \frac{2}{3}$. Especially for low q, when an object from category 1 is very rare, ICP_1 is difficult to estimate precisely. For high q, this is the case as well, because then categories 2 and 3 are rare, and so a misclassification as 2 or 3 is rare as well.

6.3.3 Effect of number of classes C

We studied $\boldsymbol{\alpha} = \left\{\frac{1}{C}\right\}_{c=1}^{C}$ for $C = 2, 3, 4$. We set $I = 300$, $K = 10$, and $\beta = 1$. The estimates are given in Table 6.1.

We see that the mean ICP_a are decreasing in the number of categories. Note that, in the case with three categories, the total probability of misclassification is $(3-1) \times 0.1225 = 0.2450$, but in the case with four categories, it is $(4-1) \times 0.0924 = 0.2772$, which is higher. The total probability of misclassification is thus increasing in C.

This means that, in order to maintain an equal probability of misclassification, the amount of concentration should increase (or: β should decrease) when there are more classes. Intuitively, this makes sense.

6.3.4 Effect of sample size I and number of repeated measurements K

We took the case $\boldsymbol{\alpha} = \left\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right\}$ and $\beta = 1$. We performed 1000 sets of simulations for the 9×7 grid defined by $K = 2$ to 18 in steps of 2 and $I = 120$ to 480 in steps of 60. The standard errors of ICP_a are given in Figure 6.3. Contour lines are derived by interpolation between the 63 grid points.

The solid contour lines in Figure 6.3 delineate points with constant standard errors of ICP_a. The dashed contour lines represent regions with constant $I \times K$, representing constant number of measurements. For a given amount of measurements $I \times K$, the standard errors can be optimized by moving along the dashed line into the region

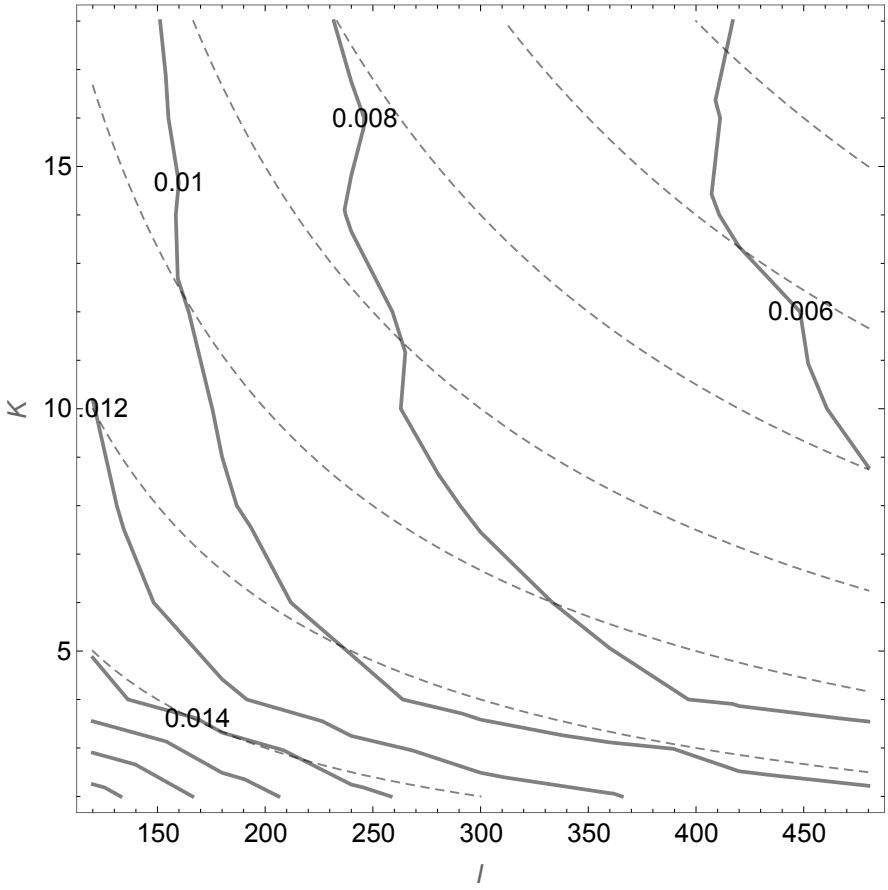


Figure 6.3: Relation between standard error, number of repeated measurements K and sample size I. Dashed lines are regions with constant $I \times K$.

with lowest standard error. We see that, in the range explored in Figure 6.3, it is optimal to lower K to 3 or 4 and let the required precision determine I.

Note that optimizing the likelihood function can be seen as fitting a distribution of response patterns to the empirical distribution of response patterns. For that reason, we recommend choosing K such that the number of possible response patterns is larger than the number of parameters in $f_{\boldsymbol{P}}$. In symbols: $\begin{pmatrix} C+K-1 \\ K \end{pmatrix} > C+1$. This means that, in the binary case, 3 repeated measurements are necessary (Van Wieringen and Van den Heuvel, 2005), and with more classes, $K = 2$ is sufficient.

6.4 Real-life case

6.4.1 Background

The case study comes from a manufacturer of decorative products. Due to a recent increase in competition, senior management has focused its strategy on product quality. This was not straightforward as no customer specifications were defined, and thus the first step was to define quality. The decision was made to distill a definition from the judgments of line operators. Specifically, the goals were to

- find out whether there are systematic differences in appraisals between operators. These differences may serve as a basis for discussion when developing a definition for quality.
- identify the operator with highest repeatability, in order to work with him or her to create a quality inspection procedure that gives consistent measurement results.

The MSA study was performed on the production line for the casings for one of the products. Management had already determined that there are 2 types of poor quality: a shortcoming in the casing that may cause the product to malfunction, or a visual shortcoming, which is almost equally relevant in a market for decorative products. We use a dataset containing measurements by 3 operators. Possible classifications are "OK" (1), "MALFUNCTION" (2) or "VISUAL" (3). After the first round of measurements, appraisers were asked to classify the products again, 2 or 3 days later, in a different, randomized order. This was to minimize the impact of the memory effect.

The sample has an important, but also realistic, complexity, which is the result of (production) time and (personnel) cost limitations. The distribution of 1s, 2s and 3s in the population of casings is, as is often the case in industry, very unbalanced, because the majority of products is conforming (1). Therefore, an extremely large sample

would be required to make sure that parts with $\tilde{X}_{ij} = 2$ or 3 are sufficiently represented for estimation of all ICP_{ja}. In particular, this means that the recommended sample size will be beyond, and specifically to the right of the range as depicted in Figure 6.3. The manufacturer chose to take a sample of 60 casings that were (subjectively) considered hard-to-judge, because it was expected that these would provide the most information when comparing operators' judgments. The sample size of 60 was chosen because this corresponded to the maximum amount of man-hours that management wanted to devote to this experiment.

We are dealing with a non-random sample of hard-to-judge parts. These are parts with $|\max_c p_{ijc} - \min_c p_{ijc}|$ relatively small. That is, especially the boundaries of the support of the Dirichlet distribution, where the $\boldsymbol{p}_{ij}$ are close to unit vectors, are underrepresented. Parts around these boundaries have a relatively low $P\left[Y_{ijk} = a | \tilde{X}_{ij} \neq a\right]$, and as the ICP_{ja} is an average of these probabilities over all products i, the estimate will turn out higher than if the sample were random. However, if appraiser $j = 1$ has $\boldsymbol{\alpha}$, β that are very much different from those of appraiser $j = 2$, this will be visible in any sample, random or non-random, in both the parameters and the ICP_{ja}. In light of the aims depicted above, this sample will thus still allow for comparing appraisers, even though the absolute values of the estimates are not representative for the entire population of products.

A more tractable approach for overcoming the challenge of an unbalanced population (with respect to X and/or $\tilde{X}$) would have been to perform conditional sampling (as explained and demonstrated in Chapters 2, 3 and 5). Conditional sampling means letting appraiser j select a mix of parts that he or she has already ($k = 0$) classified as $Y_{ij0} = 1$, $Y_{ij0} = 2$ and $Y_{ij0} = 3$, and thus obtain a sample that is more balanced. To prevent biases, the conditional probability distributions $f_{j,\boldsymbol{P}}^{Y_0=t}$ then need to be used in the likelihood (6.1).

The outcomes of the experiment are given in Table 6.2. Some of the products seemed to be straightforward to classify (product 10 was clearly malfunctioning, product 12 was clearly OK), while others were less easy to classify (for example product 21). In the next subsections, we will fit the aforementioned model to the data and give interpretations.

6.4.2 Results

The estimated parameters are in Table 6.3, and a graphical representation of the α_{jk} is given in Figure 6.4. These show that all appraisers seem to agree that this sample consists mostly of acceptable products, although in a truly random sample, this is expected to be even more clearly so. A is more strict than C is. It seems moreover that, for the nonconforming products, most problems arise because of a malfunction, not a visual shortcoming, but this result may not hold for the total population of

Case	A	B	C	Case	A	B	C	Case	A	B	C
1	1,1	1,1	1,1	21	3,3	2,2	1,3	41	1,1	1,1	1,1
2	1,1	1,1	1,1	22	2,2	2,2	2,2	42	1,3	1,1	1,1
3	2,2	2,3	1,3	23	1,1	1,1	1,1	43	1,1	1,1	1,1
4	1,3	1,1	1,1	24	2,2	2,2	2,3	44	3,3	3,3	3,3
5	2,2	2,2	1,1	25	1,2	1,1	2,2	45	2,2	1,3	1,1
6	2,2	2,2	2,2	26	2,2	2,2	1,1	46	1,1	1,1	1,3
7	2,2	2,2	1,1	27	3,3	3,3	3,3	47	3,3	3,3	1,1
8	1,1	1,1	1,1	28	1,1	1,1	1,1	48	1,1	1,1	1,1
9	3,3	1,1	1,1	29	2,2	2,2	2,2	49	1,1	1,1	1,1
10	2,2	2,2	2,2	30	2,2	2,2	2,2	50	1,1	1,1	1,1
11	1,1	1,1	1,1	31	1,1	1,1	1,1	51	2,2	2,2	1,1
12	1,1	1,1	1,1	32	1,1	1,1	1,1	52	1,1	1,1	1,1
13	1,1	1,1	1,1	33	1,1	1,1	1,1	53	1,1	1,1	1,1
14	3,3	3,3	3,3	34	1,1	1,1	1,1	54	1,1	1,1	1,1
15	1,1	1,1	1,1	35	1,1	1,1	1,1	55	1,1	1,1	1,1
16	1,2	1,1	1,1	36	1,1	1,1	2,3	56	1,1	1,1	1,1
17	2,2	2,2	1,2	37	1,1	1,1	1,1	57	1,1	1,1	1,1
18	1,1	1,1	1,1	38	1,3	1,1	1,1	58	3,3	3,3	3,3
19	2,3	1,1	1,2	39	2,2	2,2	2,3	59	1,1	1,1	1,1
20	1,1	1,1	1,1	40	1,1	1,1	1,2	60	1,1	1,1	1,1

Table 6.2: Outcome of the experiment. Categories: 1 = Product is OK, 2 = Product will lead to malfunction, 3 = Visual shortcoming.

	α_{j1}	α_{j2}	α_{j3}	β_j
A	0.5832	0.2610	0.1558	0.2597
	(0.0600)	(0.0531)	(0.0434)	(0.1159)
B	0.6687	0.2216	0.1098	0.0741
	(0.0596)	(0.0525)	(0.0393)	(0.0553)
C	0.7211	0.1530	0.1259	0.5591
	(0.0525)	(0.0413)	(0.0378)	(0.2635)

Table 6.3: Estimated parameters of the full model. Standard errors between brackets. Categories: 1 = Product is OK, 2 = Product will lead to malfunction, 3 = Visual shortcoming.

	ICP_{j1}	ICP_{j2}	ICP_{j3}
A	0.0636	0.1053	0.1183
B	0.0167	0.0364	0.0416
C	0.0809	0.1980	0.2034

Table 6.4: Probabilities of inconsistent classification. Categories: 1 = OK, 2 = MALFUNCTION, 3 = VISUAL.

casings.

The estimates for β_j show that appraiser B has a particularly high repeatability, while C has a low repeatability. The point estimate for C lies outside the confidence intervals of β of A and B.

6.4.3 Probabilities of inconsistent classification

The estimates and standard errors of the ICPs are given in Table 6.4. As reflected in the high β_j, C uniformly has the highest probability of inconsistent classification. Moreover, for all appraisers, the most inconsistent classifications are amongst products that were classified as 3. This may be explained by aesthetic value being very subjective when not objectively described. A classification "OK" is uniformly most likely to be a classification that is not inconsistent with $\tilde{X}$.

As mentioned before: these figures are not based on a random sample, so the relative sizes of the ICP are more reliable than their absolute values.

6.4.4 Hypothesis testing

We perform various likelihood ratio tests on the parameters of the model. The null hypotheses with the corresponding loglikelihood values and p-values are given in Table 6.5.

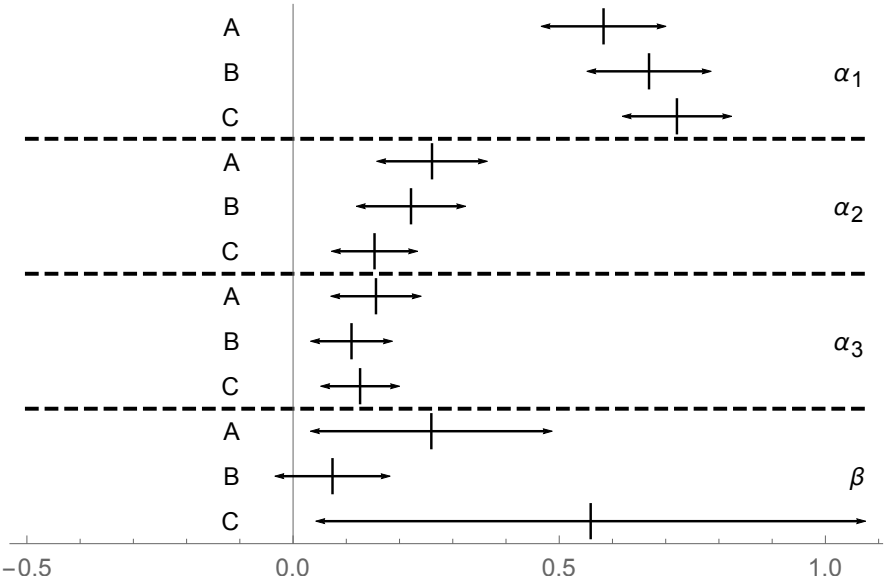


Figure 6.4: Estimated parameters of the full model. Vertical lines are point estimates, arrows mark confidence interval. Categories: 1 = OK, 2 = MALFUNCTION, 3 = VISUAL.

Type	H_0	logL	p-value
Unrestricted model		-218.91	-
Systematic	$\boldsymbol{\alpha}_A = \boldsymbol{\alpha}_B = \boldsymbol{\alpha}_C$	-220.75	0.4537
	$\boldsymbol{\alpha}_A = \boldsymbol{\alpha}_B$	-219.47	0.5739
	$\boldsymbol{\alpha}_A = \boldsymbol{\alpha}_C$	-220.54	0.1969
	$\boldsymbol{\alpha}_B = \boldsymbol{\alpha}_C$	-219.46	0.5806
Repeatability	$\beta_A = \beta_B = \beta_C$	-222.17	0.0389
	$\beta_A = \beta_B$	-220.11	0.1229
	$\beta_A = \beta_C$	-219.63	0.2327
	$\beta_B = \beta_C$	-222.16	0.0108
Exchangeability	$\alpha_A = \alpha_B = \alpha_C, \beta_A = \beta_B = \beta_C$	-223.81	0.1346
	$\boldsymbol{\alpha}_A = \boldsymbol{\alpha}_B, \beta_A = \beta_B$	-220.99	0.2469
	$\boldsymbol{\alpha}_A = \boldsymbol{\alpha}_C, \beta_A = \beta_C$	-220.85	0.2769
	$\boldsymbol{\alpha}_B = \boldsymbol{\alpha}_C, \beta_B = \beta_C$	-222.56	0.0634

Table 6.5: Tested hypotheses involving multiple appraisers. Likelihood ratios calculated with respect to unrestricted model.

Null hypothesis	Likelihood Ratio	p-value
$H_0 : f_{A,\boldsymbol{P}} = f^{\infty}_{A,\boldsymbol{P}}$	62.95	0.000
$H_0 : f_{B,\boldsymbol{P}} = f^{\infty}_{B,\boldsymbol{P}}$	78.74	0.000
$H_0 : f_{C,\boldsymbol{P}} = f^{\infty}_{C,\boldsymbol{P}}$	31.35	0.000
$H_0 : f_{A,\boldsymbol{P}} = f^{\infty}_{A,\boldsymbol{P}}, \boldsymbol{\alpha}_A = \left\{C^{-1}\right\}_{c=1}^{C}$	98.26	0.000
$H_0 : f_{B,\boldsymbol{P}} = f^{\infty}_{B,\boldsymbol{P}}, \boldsymbol{\alpha}_B = \left\{C^{-1}\right\}_{c=1}^{C}$	142.9	0.000
$H_0 : f_{C,\boldsymbol{P}} = f^{\infty}_{C,\boldsymbol{P}}, \boldsymbol{\alpha}_C = \left\{C^{-1}\right\}_{c=1}^{C}$	112.0	0.000

Table 6.6: Tested hypotheses and results.

Appraiser	G_j	p
A	0.504	0.777
B	4.545	0.103
C	4.772	0.092

Table 6.7: Results of goodness-of-fit tests.

There is evidence that there are differences in repeatability ($p = 0.0389$), which seems mostly due to differences in repeatability between appraiser B and C ($p = 0.0108$). There is no indication of any systematic difference between appraisers.

Other test results are given in Table 6.6, and are about informative vs. uninformative measurements, because as β_j goed to infinity, each object will yield the same response pattern distribution. The first three p-values show that, indeed, there is evidence for informative categorization. The latter three null hypotheses represent the special cases of random assignment ("guessing"), because of the uniform distribution of measurement outcomes. These null hypotheses are also rejected.

For testing the hypothesis that $f_{j,\boldsymbol{P}} = f^{\infty}_{j,\boldsymbol{P}}$, we used that

$$\lim_{\beta_j \to \infty} L_j \left(\{\alpha_{jc}\}_{c=1,\ldots,C}, \beta_j \right) = \log \prod_{i=1}^{I} \prod_{c=1}^{C} \alpha_{jc}^{\#\{Y_{ijk}=c\}} = \sum_{i=1}^{I} \sum_{c=1}^{C} \#\{Y_{ijk} = c\} \log \alpha_{jc}.$$

6.4.5 Diagnostics

Since there are 6 possible response patterns (2 repeated measurements over 3 classes), 5 parameters are necessary to fit all the associated frequencies (saturated model). In the proposed model 3 parameters are needed. We can perform a goodness of fit test with $df = 5 - 3 = 2$, as described in Section 6.2.3. The test results are given in Table 6.7, and the actual and fitted response patterns are given in Table 6.8.

$\boldsymbol{u}$	$E_A(\boldsymbol{u})$	$E[E_A(\boldsymbol{u})]$	G	$E_B(\boldsymbol{u})$	$E[E_B(\boldsymbol{u})]$	G	$E_C(\boldsymbol{u})$	$E[E_C(\boldsymbol{u})]$	G
$(2,0,0)$	32	32	0.0	40	39	1.6	41	39	4.2
$(1,1,0)$	3	3.8	-1.4	0	1.2	0.0	3	4.7	-2.8
$(1,0,1)$	3	2.2	1.7	1	0.6	1.0	3	3.9	-1.6
$(0,2,0)$	14	13	1.5	13	13	0.9	6	6.4	-0.8
$(0,1,1)$	1	1.0	-0.0	1	0.2	3.2	3	0.8	7.7
$(0,0,2)$	7	7.7	-1.4	5	6.2	-2.1	4	5.2	-2.1

Table 6.8: Contributions to the goodness-of-fit G-statistic.

It seems that all appraisers pass the goodness-of-fit test, although depending on the significance level, the results might be seen as weak evidence for a bad fit for C.

6.4.6 Conclusions from the case study

Although there are no significant systematic differences between appraisers, there were differences in repeatability. In particular, there was a significant difference between operators B and C, the former being more consistent.

As a result of the investigations, a discussion between the two operators was organized, with the task of formulating standard operating procedures in order to obtain consistent classifications for all operators. Moreover, parts 19, 21, 25 and 45 were classified least consistent, and were brought to a discussion with other operators and management as well. Where no agreement could be obtained, higher management made a choice.

6.5 Conclusions

This chapter proposes a method to quantify measurement error of nominal measurement systems in situations when no gold standard is available. Without a gold standard, the event of a measurement error is undetectable, and estimation of false classification probabilities $FCP_a = P[Y = a|X \neq a]$ is thus problematic (Akkerhuis et al., 2016a,b). Therefore, this chapter proposes a method to estimate the random component of measurement error only: the probabilities of inconsistent classification $ICP_a = P\left[Y = a|\tilde{X} \neq a\right]$, with $\tilde{X}$ the modal or most likely measurement outcome. We conceive the situation $X = \tilde{X}$ as an absence of systematic measurement error, which results in $ICP = FCP$. Although it may seem like an important limitation of ICP that it does not express systematic error, it is not unsurprising when compared to standard practices for numerical MSA without gold standard. In numerical measurement, it is common practice to divide measurement error into a systematic

and a random component. Systematic error is operationalized by "bias": the difference between the expected measurement outcome and the true value $E[Y] - X$. In nominal measurement, "difference" is not defined due to the measurement scale, but probabilities $P\left[\tilde{X} \neq X\right]$ offer another way of quantification. For both expressions, it is clear that its estimation is unattainable without a gold standard.

Random measurement error for numerical measurement is operationalized by the difference between measured and expected outcomes $Y - E[Y]$, where $E[Y]$ is comparable to the $\tilde{X}_i$ in the nominal case. It can be quantified by a gage R&R study without a gold standard being available. Since differences are undefined on a nominal scale, this chapter proposes to quantify random measurement error of nominal measurement by probabilities of inequality $P\left[Y = a|\tilde{X} \neq a\right]$. In that sense, the approach proposed in this chapter moves closer towards what is already common practice for numerical MSA, by acknowledging that, without a gold standard, only the random component of measurement error is estimable.

Chapter 7

Summary

This thesis is about binary classifications. Important examples are "accept"/"reject" decisions by quality inspections in industry, or "positive"/"negative" results by diagnostic tests in medicine. As these classifications aim to reflect empirical properties of products or patients, we conceive of them as measurements. The context of this thesis is industry, where quality inspections accept or reject a product, based on the customers' requirements. These inspections can be performed by persons, devices or persons operating devices, more generally called binary measurement systems.

Measurement is often not 100% reliable, because measurement errors occur. A measurement error is the event that the measurement outcome differs from the "true value". In quality inspections, a false acceptance means that the customer receives a defective product and a false rejection leads to unnecessary scrap and/or rework. When using a binary measurement system, it is therefore important to have some quantification of its reliability in terms of measurement error. To this end, the quantities *FAP* (False Acceptance Probability) and *FRP* (False Rejection Probability) are often used. These are called the probabilities of misclassification.

The true value of a product, whether or not it is truly conforming to specifications, is usually obtained by applying a "gold standard". This is an authoritative measurement procedure about which there is consensus that it gives the true value. This thesis focuses on the common situation that a gold standard is not available. This may be the case when customer specifications are not clearly defined, for example when it concerns judgments of perceptual qualities such as taste. If applying a gold standard is very time-consuming, expensive, or even destructive, it may also be considered practically unavailable.

The branch of applied statistics that is concerned with quantifying reliability of measurement systems is called "measurement system analysis" (MSA). An MSA study

consists of performing an experiment (usually involving repeated measurements), fitting a model to the data and expressing reliability of the measurements using quantifiers such as *FAP* and *FRP*. MSA for numerical measurements is a mature research topic, also for cases without gold standard. In the binary case, MSA without a gold standard attracted a lot of attention in scientific literature, but turns out to be a stubborn problem. Recent literature proposes various methods for binary MSA without gold standard, but warnings are voiced that substantial biases in estimates of *FAP* and *FRP* can occur. The situation at the start of this PhD project was that practitioners have a variety of methods to choose from for the estimation of *FAP* and *FRP*, but none of these have shown convincing proof of their performance. This thesis aims to find out what the limitations are in estimating *FAP* and *FRP* without a gold standard, and proposes a method that respects these limitations.

Chapter 2 reports on a case study, in which a tailor-made approach is presented and evaluated. It turns out that *FAP* and *FRP* cannot be estimated in this specific case. This is a result of the missing gold standard, and in particular, the unknown customer requirements. Chapter 2 ends with the suggestion that measurement error has a random component. Although this is widely known in numerical MSA, it is a new perspective on the evaluation of binary measurement systems.

Chapter 3 mathematically shows that problems regarding the estimation of *FAP* and *FRP* are not limited to the case in Chapter 2, but that these problems are actually more general: without gold standard, *FAP* and *FRP* are in general unidentifiable quantities. The suggestion that measurement error has a random component is formalized by decomposing binary measurement error in both a random and a systematic component.

In numerical MSA, the distinction between systematic and random measurement error is based on the expected value of the measurement outcome (as can be approximated without gold standard by taking the average of a series of repeated measurements). Systematic measurement error is the difference between the true value and this expected outcome, and random measurement error is dispersion of measurement outcomes around the expected outcome. Binary measurements do not have an expected value or average, but they do have a modal outcome. This is the most likely outcome: if a product receives a reject decision more often than an accept decision in repeated measurement, its modal outcome is reject. A systematic measurement error in the binary case thus means that a false outcome is given in more than half of the times. But without systematic measurement error, a measurement system still may give outcomes that randomly deviate from the modal outcome. These deviations are called random measurement errors.

The systematic component of *FAP* and *FRP* cannot be determined without gold standard. The random component does not depend on the true value of an object, but only relies on identification of the modal outcome. Therefore, repeated measure-

ments will make it possible to estimate the random component of measurement error. The random component is quantified by the inconsistent acceptance and rejection probabilities IAP and IRP: the probabilities that a measurement outcome is unequal to the modal outcome.

Despite the theoretical problems described above, methods could possibly still give useful results in practice. Therefore, Chapter 4 presents an extensive simulation study, in which various existing methods are compared. It is shown that estimators of FAP and FRP can be greatly biased if assumptions are not exactly met, because their systematic components are not identifiable. This means that currently existing methods are not useful in practice. However, the existing methods were also modified to produce estimates of IAP and IRP. One of the tested methods turns out to estimate IAP and IRP reliably across a wide range of scenarios. The model assumed in the method, however, is strongly tied to the method's original purpose of estimating FAP and FRP, and it is likely that a more parsimonious method can be developed.

Chapter 5 offers a robust technique for estimating IAP and IRP that relies less on functional form assumptions. It is an adaptive procedure that increases model complexity as long as the functional form does not have enough degrees of freedom to fit the observed data well. Each time a significance test for goodness-of-fit fails, another degree of freedom is added. The performance of the method is demonstrated in a simulation study.

Chapter 6 generalizes previous findings to the case of measurement scales with more than two classes. An example is the assignment of failure modes to failed products (there are typically more than two failures modes). The distinction between systematic and random measurement error is made explicit here as well. A method is proposed to estimate the random component. The IAP and IRP that were used in previous chapters, are generalized to $C \times (C-1)$ probabilities of inconsistent classification (ICPs).

We recommend *not* to try and estimate FAP and FRP when a gold standard is unavailable, since the thesis convincingly shows that estimates are likely to be biased, possibly substantially. Only the random components can be estimated reliably; the method presented in Chapter 5 has shown good results. The interpretation of IAP and IRP is less straightforward, but these probabilities are nevertheless useful. If IAP and IRP are substantial, this implies that the measurement system cannot consistently discriminate worse from better parts, and this diqualifies the system for an application. If IAP and IRP are reasonably good, the system's discriminative capability is sufficient for the application. If possible, it would be desirable in that case to rule out that the test results are systematically too strict of too lenient. Such would require calibration. This remains problematic without gold standard, and a rough calibration based on consensus among engineers may be the best one can do.

Bibliography

Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics*, 44:539–548.

AIAG (2003). *Measurement System Analysis: Reference Manual, 3rd ed.* Automotive Industry Action Group, Detroit, MI.

Akkerhuis, T. S. (2012). Bias and precision of error rate estimates of binary measurement systems. Master's thesis, University of Amsterdam, Amsterdam.

Akkerhuis, T. S. and De Mast, J. (2015). Quantifying the random component of measurement error of nominal measurements without a gold standard. *Quality and Reliability Engineering International*, in press.

Akkerhuis, T. S., De Mast, J., and Erdmann, T. (2016a). Estimation of the random error of binary tests using adaptive polynomials. *Submitted for publication.*

Akkerhuis, T. S., De Mast, J., and Erdmann, T. (2016b). The statistical evaluation of binary tests without gold standard: robustness of latent variable approaches. *Submitted for publication.*

Albert, P. S. and Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60:427–435.

Albert, P. S., McShane, L. M., and Shih, J. H. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard. *Biometrics*, 57:610–619.

Allen, M. J. and Yen, W. M. (1979). *Introduction to Measurement Theory.* Brooks/Cole, Monterey, CA.

Bashkansky, E. and Gadrich, T. (2013). Some statistical aspects of binary measuring systems. *Measurement*, 46:1922–1927.

BIBLIOGRAPHY

Beavers, D. P., Stamey, J. D., and Bekele, B. N. (2011). A bayesian model to assess a binary measurement system when no gold standard system is available. *Journal of Quality Technology*, 43:16–27.

Becker, M. P. and Agresti, A. (1992). Log-linear modelling of pairwise interobserver agreement on a categorical scale. *Statistics in Medicine*, 11:101–114.

Bierens, H. J. (2005). Semi-nonparametric modeling of densities on the unit interval, with applications to censored mixed proportional hazard models and ordered probability models: Identification and consistency results. *Econometric Society World Congress*, 2005.

Boyles, R. A. (2001). Gauge capability for pass-fail inspection. *Technometrics*, 43:223–229.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Danila, O., Steiner, S. H., and MacKay, R. J. (2008). Assessing a binary measurement system. *Journal of Quality Technology*, 40:310–318.

Danila, O., Steiner, S. H., and MacKay, R. J. (2010). Assessment of a binary measurement system in current use. *Journal of Quality Technology*, 42:152–164.

Danila, O., Steiner, S. H., and MacKay, R. J. (2012). Assessing a binary measurement system with varying misclassification rates using a latent class random effects model. *Journal of Quality Technology*, 44:179–191.

Danila, O., Steiner, S. H., and MacKay, R. J. (2013). Assessing a binary measurement system with varying misclassification rates when a gold standard is available. *Technometrics*, 55:335–345.

De Mast, J. (2007). Agreement and kappa-type indices. *The American Statistician*, 61:148–153.

De Mast, J., Akkerhuis, T. S., and Erdmann, T. P. (2014). The statistical evaluation of categorical measurements. *Quality Engineering*, 26:16–32.

De Mast, J., Erdmann, T. P., and Van Wieringen, W. N. (2011). Measurement system analysis for binary inspection: Continuous versus dichotomous measurands. *Journal of Quality Technology*, 42:99–112.

De Mast, J. and Van Wieringen, W. N. (2010). Modeling and evaluating repeatability and reproducibility of ordinal classifications. *Technometrics*, 52:94–106.

Erdmann, T. P. (2012). *Statistical Evaluation of Binary Measurement Systems*. PhD thesis, University of Amsterdam, Amsterdam.

Erdmann, T. P., Akkerhuis, T. S., De Mast, J., and Steiner, S. H. (2016a). The statistical evaluation of a binary test based on combined samples. *Journal of Quality Technology*, 48:54–67.

Erdmann, T. P., De Mast, J., and Warrens, M. J. (2016b). Some common errors of experimental design, interpretation and inference in agreement studies. *Statistical Methods in Medical Research*, 24:920–935.

Feinstein, A. (2002). Misguided efforts and future challenges for research on diagnostic tests. *Journal of Epidemiology Community Health*, 56:330–332.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Formann, A. K. (2003). Latent class model diagnostics - a review and some proposals. *Computational Statistics and Data Analysis*, 41:549–559.

Frederic, P. and Lad, F. (2008). Two moments of the logitnormal distribution. *Communications in statistics - Simulation and Computation*, 37:1263–1269.

Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55:363–390.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159:445–492.

Hui, S. L. and Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36:167–171.

Hunt, P. R., Friesen, M. C., Sama, S., Ryan, L., and Milton, D. (2015). Log-linear modeling of agreement among expert exposure assessors. *Annals of Occupational Hygiene*, in press.

Irwig, L. M., Bossuyt, P. M. M., Glasziou, P. P., Gatsonis, C., and Lijmer, J. G. (2002). Designing studies to ensure that estimates of test accuracy are transferable. *British Medical Journal*, 324:669–671.

ISO (1995). *Guide to the Expression of Uncertainty in Measurement, 1st ed.* International Organization for Standardization, Geneva, Switzerland.

JCGM (2008). *International Vocabulary of Metrology - Basic and General Concepts and Associated Terms (VIM), 3rd ed, ISO/IEC Guide 9912:2007*. Joint Committee for Guides in Metrology, ISO, Geneva, Switzerland.

Kallenberg, W. C. M., Oosterhoff, J., and Schriever, B. F. (1985). The number of classes in chi-squared goodness-of-fit tests. *Journal of the American Statistical Association*, 80:959–968.

Kimothi, S. K. (2002). *The Uncertainty of Measurements*. ASQ, Milwaukee.

Knottnerus, J. A., Weel, C. V., and Muris, J. W. M. (2002). Evaluation of diagnostic procedures. *British Medical Journal*, 324:477–480.

Kraemer, H. C. (1987). The methodological and statistical evaluation of medical tests: The dexamethasone suppression test in psychiatry. *Psychoneuroendocrinology*, 12:411–427.

Kraemer, H. C., Periyakoil, V. S., and Noda, A. (2004). Kappa coefficients in medical research. *Tutorials in Biostatistics*, 1:85–105.

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum, Hillsdale, NJ.

Mehrotra, S. (1992). On the implementation of a primal-dual interior point method. *SIAM Journal on Optimization*, 2:575–601.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7:308–313.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford, UK.

Qu, Y., Tan, M., and Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52:797–810.

Rice, J. R. (1975). A metalgorithm for adaptive quadrature. *Journal of the ACM*, 22:61–82.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103:677–680.

Stewart, M. B. (2004). Semi-nonparametric estimation of extended ordered probit models. *Stata Journal*, 4:27–39.

Suzuki, T., Tsutsumi, Y., and Kawamura, H. (2013). Viewpoints to characterize precision evaluation methods in binary measurements. *Measurement*, 46:3710–3714.

Tanner, M. A. and Young, M. A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, 80:175–180.

Torrance-Rynard, V. L. and Walter, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine*, 97:2157–2175.

Vacek, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41:959–968.

Van Wieringen, W. N. and De Mast, J. (2008). Measurement system analysis for binary data. *Technometrics*, 50:468–478.

Van Wieringen, W. N. and Van den Heuvel, E. R. (2005). A comparison of methods for the evaluation of binary measurement systems. *Quality Engineering*, 17:495–507.

Vardeman, S. B. and Van Valkenburg, E. S. (1999). Two-way random effects analyses and gauge R&R studies. *Technometrics*, 41:202–211.

Wallsten, T. S. (1988). Measurement theory. In Kotz, S. and Johnson, N., editors, *Encyclopedia of Statistical Sciences. Vol. 5*. Wiley, New York.

Wilrich, P. T. (2010). The determination of precision of qualitative measurement methods by interlaboratory experiments. *Accreditation and Quality Assurance*, 15:439–444.

Wolfram Research (2010). Mathematica 8.

Samenvatting

Dit proefschrift gaat over binaire classificaties. Belangrijke voorbeelden zijn kwaliteitsinspecties in de industrie die producten goedkeuren of afkeuren, en medische diagnostische tests. Aangezien deze classificaties als doel hebben om empirische eigenschappen van producten of mensen te weerspiegelen, beschouwen we ze als metingen, uitgevoerd door binaire meetsystemen. De context van dit proefschrift is de industrie, waar kwaliteitsinspecties producten toetsen aan specificaties en ze op basis hiervan goed- of afkeuren. Deze inspecties kunnen worden uitgevoerd door mensen, apparaten of mensen die apparaten bedienen.

Een meetsysteem is vaak niet 100% betrouwbaar, omdat meetfouten kunnen optreden. Een meetfout is de gebeurtenis dat de meetuitkomst afwijkt van de werkelijke waarde. Bij kwaliteitsinspecties zijn er twee soorten meetfout. Een onterechte goedkeur betekent dat de klant een defect product ontvangt, en een onterechte afkeur leidt tot onnodige verspilling van tijd en middelen. Wanneer men gebruik maakt van een binair meetsysteem, is het dan ook belangrijk om de betrouwbaarheid, in termen van meetfout, te kwantificeren. Hiervoor worden de grootheden *FAP* (*False Acceptance Probability*, de kans op onterechte goedkeur) en *FRP* (*False Rejection Probability*, de kans op onterechte afkeur) vaak gebruikt. Deze heten misclassificatiekansen.

De eerder genoemde werkelijke waarde van een product, namelijk of deze in realiteit wel of niet voldoet aan de specificaties, wordt meestal verkregen door een gouden standaard toe te passen. Dit is een meetprocedure waarover experts het eens zijn dat deze tot de werkelijke waarde leidt. In dit proefschrift onderzoeken we de veelvoorkomende situatie dat deze standaard niet beschikbaar is. Dit gebeurt onder andere wanneer de specificaties niet precies zijn, bijvoorbeeld wanneer het om smaak gaat. Als een gouden standaard te tijdrovend, duur, of zelfs destructief is, wordt deze ook vaak opgevat als praktisch niet beschikbaar.

De tak van de toegepaste statistiek die zich bezighoudt met het kwantificeren van de betrouwbaarheid van meetsystemen heet meetsysteemanalyse (MSA, *Measurement System Analysis*). Een MSA-studie bestaat uit het uitvoeren van een experiment

(vaak met herhaalde metingen), het fitten van een model op de data en het gebruiken van dit model om meetfout in grootheden zoals *FAP* en *FRP* uit te drukken.

MSA voor numerieke metingen is een volwassen onderzoeksgebied, ook voor de gevallen zonder gouden standaard. In het binaire geval heeft MSA zonder gouden standaard weliswaar veel aandacht gekregen in de literatuur, maar het lijkt een hardnekkig probleem. Er zijn veel methoden voorgesteld voor binaire MSA zonder gouden standaard, maar er wordt gewaarschuwd dat deze kunnen leiden tot aanzienlijke schattingsfouten in *FAP* en *FRP*. De situatie aan het begin van dit PhD-project was dat praktijkbeoefenaars konden kiezen uit een groot aanbod aan methoden, maar dat voor geen van de methoden overtuigend bewijs was geleverd dat de schattingsfout binnen de perken blijft. Dit proefschrift bepaalt wat precies de beperkingen zijn voor het schatten van *FAP* en *FRP* als gevolg van het ontbreken van een gouden standaard, en stelt een methode voor die binnen deze beperkingen zinvolle resultaten geeft.

Hoofdstuk 2 doet verslag van een casestudie, waar een maatwerkmodel wordt gepresenteerd en geëvalueerd. Het blijkt dat *FAP* en *FRP* in dit specifieke geval niet geschat kunnen worden. Dit komt doordat een gouden standaard ontbreekt, en in het bijzonder doordat de specificaties onbekend zijn. Het hoofdstuk sluit af met de suggestie dat binaire meetfout een toevallige component heeft. Hoewel dit algemeen bekend is voor numerieke metingen, is het een nieuw perspectief op binaire metingen.

Hoofdstuk 3 laat wiskundig zien dat de problemen met het schatten van *FAP* en *FRP* in hoofdstuk 2 niet alleen gelden voor die specifieke casus, maar dat deze problemen algemener zijn: zonder gouden standaard, zijn *FAP* en *FRP* typisch niet identificeerbaar. De suggestie uit het voorgaande hoofdstuk dat binaire meetfout een toevallige component heeft, wordt concreet gemaakt door de binaire meetfout op te splitsen in een systematische en een toevallige component.

In numerieke MSA is het verschil tussen systematische en toevallige meetfout gebaseerd op de verwachtingswaarde van de meetuitkomst; deze kan zonder gouden standaard worden benaderd door het gemiddelde te nemen van een reeks herhaalde metingen. Systematische meetfout is het verschil tussen de werkelijke waarde en deze verwachtingswaarde, en toevallige meetfout is de spreiding van de meetuitkomsten rondom de verwachtingswaarde. Binaire metingen daarentegen hebben geen betekenisvolle verwachtingswaarde of gemiddelde, maar wel een modale uitkomst. Dit is de meest waarschijnlijke uitkomst: als een product onder herhaalde metingen vaker wordt afgekeurd dan goedgekeurd, dan is de modale uitkomst een afkeur. Een systematische meetfout in het binaire geval betekent dus dat een verkeerde uitkomst wordt gegeven in meer dan de helft van de gevallen. Zonder systematische meetfout kan een meetsysteem echter nog steeds uitkomsten geven die toevallig afwijken van de modale meetuitkomst. Deze afwijkingen worden toevallige meetfout genoemd.

Onze analyse laat zien dat de systematische component van *FAP* en *FRP*

niet bepaald kan worden zonder gouden standaard. Dit verklaart grotendeels de hardnekkigheid van de problemen in het vinden van een betrouwbare methode om *FAP* en *FRP* te schatten. De toevallige component hangt niet af van de werkelijke waarde, maar enkel van de modale uitkomst. Daarom kan men deze wel bepalen zonder een gouden standaard toe te passen. De toevallige component wordt uitgedrukt in kansen van inconsistente classificatie *IAP* en *IRP*: de kansen dat de meetuitkomst inconsistent is met de modale uitkomst.

Ondanks de theoretische problemen die in hoofdstuk 3 worden beschreven, is het mogelijk dat bestaande methoden praktisch bruikbare resultaten geven, al zijn ze niet 100% nauwkeurig. Hoofdstuk 4 onderzoekt of dit zo is met een uitgebreide simulatiestudie waarin bestaande methoden worden getoetst en vergeleken. Het hoofdstuk laat zien dat schatters voor *FAP* en *FRP* tot substantiële misschattingen kunnen leiden, omdat hun systematische componenten niet bepaald kunnen worden. Dit betekent dat de bestaande methoden amper bruikbaar zijn in de praktijk. Echter, vervolgens zijn de methoden aangepast om schattingen van *IAP* en *IRP* te geven. Bijna alle methoden lijken hiervoor vrijheidsgraden tekort te schieten, maar de methode met de meeste parameters geeft betrouwbare schattingen onder een grote diversiteit aan scenario's. Het model dat hieraan ten grondslag ligt is echter gericht op het bepalen van *FAP* en *FRP*, en het is waarschijnlijk dat een doeltreffendere methode ontwikkeld kan worden.

Hoofdstuk 5 presenteert een robuuste methode voor het schatten van *IAP* en *IRP*. Het is een adaptieve procedure die stapsgewijs vrijheidsgraden aan het model toevoegt tot deze de data goed fit. De betrouwbaarheid van de methode wordt gedemonstreerd met een simulatiestudie.

In hoofdstuk 6 wordt het algemenere geval van nominale metingen behandeld. Bij nominale metingen zijn er $C > 2$ uitkomsten mogelijk. Bijvoorbeeld wanneer een faalwijze moet worden toegewezen aan een afgekeurd product (typisch zijn er meer dan twee faalwijzen). Het onderscheid tussen systematische en toevallige meetfout wordt ook hier expliciet gemaakt. Een methode wordt voorgesteld om de toevallige component van meetfout te schatten. De *IAP* en *IRP* uit de vorige hoofdstukken worden gegeneraliseerd naar de $C \times (C-1)$ kansen op inconsistente classificatie (de *ICP*'s).

De bijdrage van dit proefschrift kan worden samengevat in het volgende advies. We raden af om te proberen om *FAP* en *FRP* te schatten wanneer een gouden standaard niet voorhanden is: hoofdstukken 3 en 4 laten duidelijk zien dat misschattingen waarschijnlijk en substantieel zijn. Alleen de toevallige component, gekwantificeerd als kansen op inconsistente classificatie *IAP* en *IRP*, kan betrouwbaar worden geschat – de methode in hoofdstuk 5 laat goede resultaten zien. Als deze kansen groot zijn, dan betekent dit dat het meetsysteem niet consistent de slechtere van de betere producten kan onderscheiden, en dus dat het meetsysteem niet geschikt is. Als ze

echter laag zijn, dan is het onderscheidend vermogen van het meetsysteem wel goed. Het verdient dan aanbeveling om, waar mogelijk, systematische meetfout uit te sluiten met een calibratiestudie. Dit blijft problematisch zonder gouden standaard, en een grove calibratie die gebaseerd is op consensus onder experts is dan het beste wat men kan doen.

Curriculum vitae

Thomas Simon Akkerhuis (1988) went to secondary school at the Gymnasium Celeanum in Zwolle. His core courses were biology, chemistry, mathematics and physics, and he followed elective courses in economics, Latin and philosophy. Together with two classmates, he won a silver medal in the European Union Science Olympiad in his fifth year. After graduation, he moved to Amsterdam.

In 2011 he received his BSc (*cum laude*) in Econometrics and Operations Research at the University of Amsterdam (UvA). In 2012, he obtained an MSc degree in Econometrics. He wrote his thesis "Bias and precision of error rate estimates of binary measurement systems" during an internship at the Institute for Business and Industrial Statistics (IBIS UvA). After obtaining his MSc degree, he started working there. Besides doing scientific research, he spent his time on training professionals and giving lectures to students at the UvA.

Under supervision of his promotor Prof. Jeroen de Mast, and support of Dr. Tashi Erdmann who was finishing his thesis on this subject at the time, Thomas continued the research he started with in his MSc thesis.

With the support of, among others, Prof. Ronald Does, director of IBIS UvA, Thomas grew to be a consultant for professionals who want to improve their business processes. This means delivering Lean Six Sigma trainings, and giving clients advice on how to apply Lean Six Sigma in the real world. These clients originate from many different sectors, such as financial services, government, healthcare, and industry. Together with these clients, he wrote articles about succesful improvement projects.

At the Amsterdam Business School, Thomas has given courses to students about process improvement, statistics and mathematics. Many of these courses were for executive students, but he taught to regular BSc- and MSc-degree students as well.

Thomas is treasurer of a large home-owners coorperation. He likes to spend his free time with friends and music, preferably outside in a 25 $^{\circ}C$ (or higher) climate.

Acknowledgements

In these last pages I want to express my gratitude to all who have directly, or indirectly, contributed to this thesis.

Most importantly, I thank Jeroen de Mast and Ronald Does. Jeroen has been a great supervisor for his excellent coaching skills, endless patience and helpful advice (not only regarding the research). The amount of time he devoted to the research subject itself is more than one can reasonably expect from a promotor; his enthusiasm has been contagious. Ronald is the director of IBIS UvA, and he has given me the opportunity to combine my scientific research with training, consultancy and lecturing, which were indispensable activities to recharge my batteries after long days of performing simulations and rewriting papers. His teaching and management style have been a great inspiration. Jeroen and Ronald, thank you both very much.

Other very important contributors to this thesis are Tashi Erdmann and Michel Mandjes. Tashi was finishing his thesis on binary measurement system analysis when I joined IBIS UvA, and made me familiar with the subject and the art of academic writing. Moreover, he provided useful feedback on my first months of consulting and made me feel welcome. Michel was the second supervisor and had a great role in the final stages of writing the thesis. Rob Cuperus, Stefan Steiner and Mark Vertogen have also greatly contributed by providing feedback and real-life data.

I would like to thank my colleagues Atie Buisman, Marit Schoonhoven, Merve Güvendik and Rob Goedhart for their support and the great working atmosphere. Speaking of working atmosphere: the people working at the Institute of Executive Programmes also deserve many thanks for the amusing talks, the table football matches, and the Friday afternoon drinks that incidentally take place on other days as well. Special thanks go to my colleagues Alex and Inez, for their support, constructive brainstorming sessions and friendship at, and outside of, work. Finally, without the efficient and pleasant collaboration with Sander Bot during my BSc and MSc, I might have been in my thirties by the time I finished this.

A week consists of roughly 50 hours sleeping, 40 hours working and 40 hours of "overhead", which leaves me with 38 hours every week to work on my work-life

balance. Luckily, I am surrounded by people that helped me keep that balance. Those people made these years of working on this thesis a great period. Foremost, I am indebted to my parents Stella and Gerrie for all the ways they have supported me. Furthermore, I owe a lot to my dear friends Channah, Desiré, Enzo, Funda, Jurre, Laura, Mason, Melanie, Ruben, Sander van Ommeren, Sander Wichers, Sebastiaan, Sophie, Sora, Stijn, Zen and Zillah – and those friends who also happen to be my family: Anne (my sister), Nicolien and Suzanne (my cousins). You all made me look forward to my lunches, evenings and/or weekends, and sometimes even made it hard to go back to work after that, which I consider to be a healthy thing.

The defense of this thesis is planned for the 16th of September (2016). It is one of the rare days in life in which all important things come together: years of research, my family and my friends. I thank my sister Anne and my cousin Suzanne, my paranymphs, very much for all their efforts in making this a great day.

Thomas Akkerhuis
July 2016