MASTER

A field investigation in the lithographic industry

do printing machine operators outperform a formula on predictions of throughput time for a set of print jobs?

van den Bogaard, B.L.H.

Award date:
2010

Eindhoven, 22 June 2010

A Field Investigation in the Lithographic ndustry:

Do printing machine operators outperform a formula on predictions of throughput time for a set of print jobs?

by Bart van den Bogaard

Identity number 0618133

in partial fulfilment of the requirements for the degree of

Master of Science

in Human-Technology Interaction

Supervisors:

Prof. Dr. C.C.P. (Chris) Snijders

Dr. J.R.C. (Jaap) Ham

Abstract

The literature on ‘clinical versus statistical’ predictions has shown that experts know much, but predict poorly compared to simple statistical formulas. A lack of feedback often is attributed to this phenomenon. Without being confronted with the outcomes of one’s prediction, it is impossible to learn and consequently improve performance. Moreover, the majority of studies on which this conclusion is based on, researched extensively the psychological and medical realm. In these domains prediction often involves human behavior. Counterintuitively, this is regarded as one of the most complex entities and somehow therefore it is logical that a statistical formula is more accurate.

In this study, a field investigation in the lithographic industry was conducted. Part of a printing machine operator’s work involves making predictions on the total throughput time for a set of print jobs. As this type of prediction is well-known and adequate feedback on the outcomes of one’s prediction is common, they were reputed as the perfect candidates to outperform a statistical formula. However, a comparison between the operators’ predictions and those as generated by a statistical formula yielded a result that favored the latter one. A more thorough analysis of the data even showed that neither the most experienced nor the most confident operators were able to outperform a statistical formula. Interestingly, under highly specific circumstances it was found that operators performed similar to a four-predictor linear model but that they were still outperformed by a heuristic that is commonly used in this industry. Based on an extensive analysis it is therefore inescapable to conclude that despite the massive opportunity to learn from one’s own mistakes, printing machine operators are unable to outperform a statistical formula when predicting the total throughput time for a set of print jobs. More general, this implies the superiority of statistical prediction even for an environment where feedback on the outcome of one’s prediction is common, task characteristics are regarded as simple and where similar conditions arise from day-to-day.

Preface

This thesis describes the graduation project of Bart van den Bogaard for the Master in Human-Technology Interaction at Eindhoven University of Technology. Over the last 14 months, I have worked intensively and with ups-and-downs on this project. People from the design department of Océ Technologies, a renowned printer manufacturer, proposed the initial idea for this research. However, this thesis describes the work from an additional study that was conducted independently. The reason for this is that the results obtained at the time my contract ended, did not yield a proper answer on my research question. In retrospect, I admit that this was the right decision to acquire valuable skills that I would not have learned otherwise. However, I have to say that there were hard times to stay motivated and focused in order to bring this project to a successful ending. To express my feelings of this achievement I would like to refer to a beautiful quotation of Nelson Mandela:

"I have walked that long road to freedom. I have tried not to falter; I have made missteps along the way. But I have discovered the secret that after climbing a great hill, one only finds that there are many more hills to climb. I have taken a moment here to rest, to steal a view of the glorious vista that surrounds me, to look back on the distance I have come. But I can rest only for a moment, for with freedom comes responsibilities, and I dare not linger, for my long walk is not yet ended" (Mandela, 1995, p. 751)

All this would be impossible, however, without the support of the people that believed in me and inspired me to continue with this project. One of those people I would like to thank is my first supervisor, Chris Snijders. First, I gratefully appreciate his willingness to pick up this project after it was already started and for providing the academic building blocks to make it an interesting project. Moreover, *"Look before you leap"* perhaps best describes the way he tried to keep me away from the pitfalls I would definitely stepped into (and yet still did sometimes), given my hands-on personality. In addition, I would like to thank my second supervisor, Jaap Ham, for reviewing my drafts and answering my questions.

Secondly, thanks go out to Jennek Geels and Bas Hermus from Océ Technologies for coaching me during the first six months of this project. Their involvement shed a light on the interests of different stakeholders and the difficulty of dealing with these constraints optimally, while leveraging everything I have learned to maximize my own satisfaction with this project.

Thirdly, thanks to Han Tanis for giving me the opportunity to get in contact with the manager of the printing department of a large insurance company, Peter Roos. Without their help, I was unable to build up a good relation with operators. The enthusiasm and friendliness of these people allowed me to try out different ideas. Their hospitality meant a lot to me, really.

Furthermore, gratitude for all those people that surrounded me during my thesis and were there to listen, motivate me, or just to hang out with. More in particular, I thank Roel Hendrikx for reviewing my drafts. Finally yet important, this would all not be possible without the support of my parents that gave me the opportunity to study and to live my dreams. Thank you!

Bart van den Bogaard,

Eindhoven, June 2010.

Table of contents

Introduction

In the lithographic industry, a large number of decisions are made daily and the outcome is often known shortly thereafter. Printing machine operators questions and make decisions about the print process such as: '*is this print job going to be on time?'; 'how late will this job be finished?'; 'can we take this job on in the light of our other commitments?'; 'how are we going to cope nowthat this machine has broken down?'* (Button & Sharrock, 1997). In order to answer these questions, one is required to have an overview of the work that is scheduled and be able to estimate the amount of time it takes to finish a set of print jobs. A field study, conducted in the light of this research, showed that tools used for planning work today largely are traditional, meaning that calculations are mostly performed *'in-the-head'* or using *'pen-and-paper'.* This suggests that the outcome of a prediction strongly depends on human expertise.

As a matter of fact, the literature on 'clinical versus statistical' prediction suggests that humans are typically not good at making accurate predictions, and are often outperformed by relatively simple statistical formulas (Grove & Meehl, 1996). The consensus about why that is, is that for humans frequent, and direct feedback is often lacking. Without systematic feedback it is hard, and sometimes even impossible to learn from one's own mistakes and judgments, no matter how experienced one is (Snijders, Tazelaar, & Batenburg, 2003; Tazelaar & Snijders, 2004). This implies that humans might still have the upper hand in tasks in which frequent and direct feedback is the norm.

As part of a printing machine operators' job involves making predictions on the total time required for printing a large number of jobs. In this thesis, it is proposed they are the perfect candidates to test whether humans are able to outperform a formula. Hence, an experimental field study was conducted to investigate if attending operators can accurately predict *(clinical prediction)* the duration of a set of print jobs and compared their performance against that of a simple formula, including a common used heuristic *(statistical prediction).*

From a pragmatic point of view, this research is interesting because knowing throughput - per time - is a prerequisite for matching workload to capacity and consequently having an 'up-to-date' planning (Dexter, Traub, & Qian, 1998). Inaccurate predictions not only decreases capacity utilization, improved scheduling reduces costs by reducing daily hours of underutilized time (work done earlier than predicted) and/or over-utilized time (work done later than predicted). Further, it reduces pressure on the work of operators and the effects also extend to customers by reducing unnecessary waiting time and consequently increasing satisfaction (Richins & Holmes, 1998).

The results of a linear regression analysis are presented in this thesis. The formula that is put forward by this method, generates a prediction on throughput time for a set of print jobs and has to compete against experienced machine printing operators. This method has already been shown to be successful in a number of studies on 'clinical versus statistical' prediction (Grove et al., 2000). However, none of the studies has rigorously addressed the ability of the people working in an environment where adequate feedback is common. Therefore, it is of particular interest to test whether printing machine operators are an exception to the general conclusion.

This thesis is organized as follow: the first chapter the literature on 'clinical versus statistical' prediction is briefly reviewed followed by a brief background on the lithographic industry. Based on this, the problem statement and research question is formulated. Subsequently, the second chapter describes what methods were used to yield an answer on this question. The third chapter describes what the role of printer data was throughput the entire study, and in the fourth chapter the results are presented. Lastly, in chapter five, the general conclusion and discussion is presented.

1. Background: Theory and Lithographic industry

The scientific domain that is associated with the comparison of predictions from humans to those generated by a statistical model, is often referred to as 'clinical versus statistical' prediction. However, before providing a background on the general findings, it is necessary to define several key terms.

First, the definition of *clinical prediction* refers to a situation in which a decision-maker is confronted with data on a number of dimensions and subsequently needs to make a prediction (Snijders et al., 2003). Although this form of decision-making is an integral part of modern society, the subject of research is mostly associated with the psychological and medical literature, for example, a behavioral psychologist who has to assess a patient on a behavioral disorder, or a medical specialist who prescribes the best possible treatment for his patient. Although experts' decisions have a substantial share in the consequence of outcomes, the literature suggests that experts know much but that it becomes impossible to make accurate predictions when adequate feedback is lacking (cf. Galtung, 2003).

Secondly, the definition of *mechanical or statistical prediction* refers to a situation where based on data from several dimensions, a formula is used to make a certain prediction (Snijders et al., 2003). The method is applied to the same situations as mentioned earlier, only now it is not the human 'expert' (i.e. psychologist, medical specialist), but instead the decision is made by a formula.

Lastly, *clinical versus statistical prediction* is the definition associated with a research domain in which the performance of a human decision-maker is compared to that of a mechanical or statistical predictor (i.e. formula). A robust conclusion from a meta-analysis of 136 studies has learned us that humans are typically not good at making these kinds of decisions and that the statistical method is equal or superior to informal clinical judgments (Grove et al., 2000). Note that this conclusion is based on a review study that included studies in the realm of psychology and medicine, in which studies that attempted to predict non-human outcomes (e.g. horse races, weather, and stock market prices) were deliberately excluded. However, it seems that also for these domains the statistical approach is superior to a human decision-maker. For example, Snijders et al. (2003); Tazelaar & Snijders (2004) has shown us that the above stated conclusion also applies to other domains as for example, the field of purchasing management. They investigated how well purchasing managers are able to predict the likelihood of problems for a given purchasing transaction. Results showed that managers certainly do not outperform freshman students and that their performance gets even worse with growing experience. Alternatively, they established a simple formula (based on a number of attributes) and compared its performance against that of the managers. Surprisingly, the formula outperformed the managers. However, this finding was partly attributed to a lack of direct and immediate feedback and therefore managers are not able to learn from their mistakes and to improve upon their predictions, it was argued.

A general observation in the field of 'clinical versus statistical' prediction is that feedback, indeed, improves individuals' performance. However, it only moves them closer to the more experienced ones but does not enable the latter to outperform a statistical formula (Grove & Meehl, 1996). Accordingly, a decision-maker benefits from outcome feedback in a way that he or she is able to learn from experience and reaches a higher accuracy rate (Graham, 1971; Schroeder, 1972). Moreover, based on empirical evidence, Ericsson (2004) concludes that ideal conditions for improving expert performance are activities such as detailed and immediate feedback on performance. In addition, he states that initial proficiency in every day and professional skills is attained within weeks and months, but that developments to a very high level of achievement appears to require many years or of even decades of experience. Likewise, the theory of deliberate practices states that a maximum level of performance for individuals in a given domain is not attainted automatically as function of extended experience, but that adequate feedback is a key factor to improve performance (Ericsson, Krampe, & Tesch-Romer, 1993. Lastly, it is argued that experts in most domains attain their highest level of performance a decade or more after physical maturation points to the importance of extensive preparation (Ericsson & Lehmann, 1996). Taken together, these findings suggest that people who have been working for a prolonged period (i.e. more than 10 years) in a domain and are used to get adequate feedback on the outcome of their predictions are the perfect candidates to compete against a statistical formula.

Moreover, Shanteau (1987) distinguishes between different types of domains in which competent performance has been observed. Generally, experts seem to outperform novices in domains where similar conditions arise from day-to-day and involve static tasks (or objects). One of such areas where high level of expertise is common and processes are highly standardized is the field of surgery. According to Ehrenwerth et al. (2006), surgeons have developed their professional skills and knowledge through *feedback* and *reflection* over an extensive period of hands-on practice. Interestingly, Wright et al. (1996) conducted a study in which they compared surgeons' time predictions for elective cases with those of a commercial scheduling software package. Results showed that surgeons provided more accurate time predictions than the software, whereas a linear regression model that incorporated the surgeon's sophisticated knowledge with the output from the standard scheduling software achieved optimal performance. Furthermore, they argued that further improvements would be likely if the information system could provide timely historical data and feedback.

Looking more closely at the domains in which statistical models seem to perform well, Snijders et al. (2003) describes three important characteristics that often are in common. First, experience and intuition are considered to play an important role. Secondly, decisions or predictions involve the incorporation of a relatively large number of dimensions (i.e. more than 5). Lastly, decisions or predictions are made in an environment in which it often not immediately obvious which attributes should be included, the weight they should get assigned or include attributes that cannot be measured exactly (as often is the case with human behavior).

Taken together, the findings presented so far imply that people working in domains where tasks are *highly repetitive, adequate outcome feedback is available* and a *limited number of dimensions* should be incorporated for a prediction, are the perfect candidates to outperform a statistical formula. One of such areas that satisfy these criteria is the lithography industry. Here, printing machine operators make predictions about the total time required for printing a large number of jobs and get feedback shortly thereafter. The duration listed for a given set of print jobs includes an approximation of the time required for how long jobs are running, what allowances for preparation, downtime and delays, and the anticipation of the arrival of new, unexpected or emergency work (cf. Button & Sharrock, 2002). The operator usually supplies this or, alternatively, a 'fixed' time is added by the planner or manager. In the upcoming paragraph, a more detailed description of different print processes is provided as well as the attributes that are important to consider when making a prediction on the total time required for printing a large number of jobs.

1.1 Lithographic industry

The print industry can be classified into two categories: *on-demand printing* and *high-volume printing*, based on the activity they perform (Rai, Duke, Lowe, Quan-Trotter, & Scheermesser, 2009). *On-demand printing* is a customer-oriented, small-scale process in which the focus is on producing several copies of identical documents such as handouts, manuals, and instruction charts. Job sizes are relatively small and often require additional finishing options, such as punching, lamination, or cutting. *High-volume printing* is a business-oriented process that is characterized by the production of a large number of documents such as invoices and policies. Submission takes place electronically and orders involve various job sizes that often require printing on different paper.

In this study, we focus on high-volume printing. The reason for this is that the process is automated for the largest part and therefore least subjective to human error. Most importantly, it generates data that is used to establish a statistical formula as well as for the design of an experimental task. Part of operators' work in this industry, is to make a prediction on the total time that is required for printing a large number of jobs, also referred to as *total throughput time*. This term is made up of the actual time a printer is running *(machine time),* plus the time that the printer status is idle *(downtime)*. In order to get a better understanding of the operators' work and to get to know which attributes contribute to a prediction of throughput time, an explorative study was conducted. Four operators were interviewed and observed in their work place for a three-week period. Results showed that the following attributes needed to be incorporated:

- Total number of impressions
- Total number of jobs
- Proportion one-sided versus two-sided printed sheets
- The different types of paper (and weights)
- The number of sets (every client has his own paper, therefore extra preparation is required)
- Interpersonal differences between operators

Each attribute influences print capacity to some extent, expressed in the total number of impressions per hour. Note that an impression refers to the number of prints and has an indirect relation with the number of sheets (i.e. two impressions can be printed on a single sheet). Theoretically, an optimal prediction is achieved by assigning a weight to each attribute and determine its impact on the nominal print capacity. In practice, however, this is usually expressed as a number of impressions that is subtracted from the printer's hourly capacity. For example, a printer has a nominal hourly capacity of 17.280 impressions. However, an operator might predict the total throughput time based on an hourly capacity of 10.000 impressions instead.

Moreover, Button and Sharrock (2002) state that an extensive familiarity with the print techniques and practices acquired through experience makes it easier to assess the impact of the attributes on print capacity. They even claim that operators can develop intensely personalized 'relationships' with their machines and can come to be recognized as having the capacity to control it in a more refined and intricate way, and certainly more than anyone less familiar with it. This suggests that operators are in the ideal position to learn the dynamics of the printer so that they intuitively ‘know’ when to replenish paper trays, inks and its hourly capacity by heart.

1.2 Problem statement & Research question

The consensus in the field of ‘clinical versus statistical’ predictions is that people often are outperformed by simple statistical formulas. The main objection to this conclusion is that adequate feedback is often lacking and therefore an individual is unable to learn from his or her mistakes. Based on the results brought up so far, the lithographic industry seems to be an environment where adequate forms of feedback are common and task are highly repetitive.

As said before, one of the tasks from an operator in printing involves making predictions about the total throughput time for set of print jobs (Button & Sharrock, 2002). For this research, one can think of an operator who daily makes a prediction on total throughput time for a large number of jobs, let say 110. Based on a number of attributes for a set of print jobs he or she determines the print capacity that is appropriate, incorporates a factor for downtime, and consequently predicts the total throughput time. If everything progresses according to ‘plan’, work is finished by the time that was predicted. If not, the operator faces the consequences of it. This usually results in overtime work or when crossing important deadlines in a serious warning from the manager. Given that there a many opportunities for learning and similar conditions arise from day-to-day, printing machine operators might be the perfect candidates to test whether they are able to outperform a statistical formula. This leads to the following research question:

> *[RQ 1] Do humans outperform statistical models when estimating the total throughput time of print jobs in the lithography industry?*

If operators outperform statistical models here, this would provide one of the rare cases where human expertise and experience beats cold statistical averages. However, in order to answer this question, it needs to be determined first how to accurately an operator’s ability to predict throughput time for a set of print jobs. However, before elaborating on this topic we first discuss the relevant methods that are used for answering this research questions in the next section.

2. Methodology

This chapter describes the different steps that were undertaken in order to address the research question. To that extent, first the relevance of conducting an explorative study in the lithographic industry is laid down followed by the importance of collecting printer data. In the next section, the relevant aspects of the clinical condition are described, followed by those from the statistical condition. In the last section it is explained how the two conditions are compared.

2.1 Explorative study

At the very start of this research, an explorative study was conducted. This was done to get a better understanding of the lithographic industry and to get to know the people working there. More in particular, printing machine operators performed a central role in this study. First, because they are the ones who are supposed to know what attributes are relevant when making a prediction on total throughput time. Secondly, because they know how an experimental task should look so that it is understandable and generalizable to other companies. The techniques that were used to acquire this information were:

- *Open-interviews:* helped in developing a deeper understanding how a prediction on total throughput time is made.
- *Observations:* provided insights about the methods that were used for making a prediction on total throughput time (e.g. calculator or pen-and-paper).

2.2 Collection of printer data

The first and foremost aspect of this research is the collection of printer data. A prerequisite for this is that the data comes from a high-volume print process. Furthermore, it needs to contain a detailed record about the duration to print a job as well as the document characteristics.

2.3 Clinical prediction

The aim of the clinical prediction is to assess participants on their ability to predict the throughput time for a set of print jobs. In the subsections the characteristics of the target group, the study-design and the experimental task are described.

2.3.1 Target group

Three groups of people participated in this study. In this way, we are able to compare what the effect of frequent and direct feedback will be. Participants are selected based on to what extent they are used to make predictions on throughput time. For every group we targeted on 32 participants. Note that the groups are listed according to their familiarity with high-volume printing.

- *Students* – participants studied at the Eindhoven University of Technology and were recruited during lunch breaks. This group served as a control group.
- *Managers* – participants had to deal with issues on planning work regularly. As managers' decision are often of large impact, it is of particular to test how well they are able to make a prediction on throughput time.
- *Operators* – All participants we required to be highly familiar with A4 document printing. To overcome that operators experienced difficulties predicting throughput times caused by a lack of

expertise, we targeted on workers that were in their 30s and 40s and had a minimum of 10 years of work experience.

2.3.2 Field investigation

Most studies on 'clinical versus statistical' prediction were conducted in controlled settings (Grove et al., 2000). However, it is argued that such a setting introduces an extra constraint because the experimental task is often much more abstract than what participants are used too (Tazelaar & Snijders, 2004, p. 220). To overcome that people feel detached from their work and consequently perform worse, a field investigation was conducted. This means that operators and managers were tested in their workplace whereas students were tested on campus. Consequently, it was decided to conduct vignette experiment that was generalizable to a large number of companies and easy to administer. The details of this method are presented in the next section.

2.3.3 Vignette Experiment

The method to test participants on their ability to make a prediction, is a vignette experiment (Snijders et al., 2003). A vignette provides a self-contained overview of a set of print jobs for which a participant needs to make a prediction on the total throughput time. To overcome that the vignettes are too abstract, a field investigation was conducted. The aim of this was to get feedback on the layout of the vignettes so that they looked realistic and were generalizable to other companies. The final concept was evaluated during a pilot study in which three operators and three students participated.

2.3.4 Procedure of the vignette experiment

The aim of the vignette experiment was to assess participants on their ability to make a prediction on throughput time and was completely pen-and-paper based. At the start of the experiment, participants were informed about the purpose of the study, that none of the information would be disclosed, and that they were free to use any method, as long as they would not forget to mention it on the last page. The task was made up of the following parts:

- Demographics
- Task description
- Vignette experiment
- A section to write down how participants arrived at their answers (e.g. calculator)

2.3.4.1 Materials required for the vignette experiment

In order to complete the vignette experiment, the following materials were needed:

- A pencil or a pen
- Calculator
- The vignettes (including the demographics and the information about the printer and process)

2.3.5 Involvement of companies

In order to get the required number of operators and managers involved in the study, companies had to be found that satisfied the following criteria:

- The core business needed to be high-volume printing (similar to where the printer data was collected)
- The majority of documents printed need to be standard (i.e. A4, 80 grams)
- All companies need to make use of a similar type of printer (i.e. Xerox Nuvera 288)

A Xerox account manager offered help and provided a list of 14 companies that satisfied these criteria. Unfortunately, for various reasons only nine companies took part in this study. The first and foremost reason was that most companies were too busy for the time of the year and had difficulties to free up time. As reward for cooperating with us, all companies received a copy of this report (which is also available in the library of the Eindhoven University of Technology).

2.4 Statistical prediction

The aim of the statistical prediction was to calculate a formula that generates a prediction for a set of print jobs. In the subsection the kinds of formulas that were included are described.

2.4.1 Multiple linear regressions analysis

When calculating a statistical formula to generate a prediction on throughput time, it is highly recommend to use (weighted) linear regression because this method yields the best results (Grove & Meehl, 1996, p. 6). This method is used to predict an outcome variable Y (i.e. total throughput time) out of a number of attributes, using the equation of a straight line. Given that we have collected several values for these variables, using the printer data, the unknown parameters (coefficients) for each X value can be calculated. This is done by fitting a model (straight line) through the data for which the sum of the squared differences between line and the actual data points is minimized, also referred to as the method of least squares (Field, 2005).

For this study, three formulas were calculated that each varied in predictive power and accuracy. A reason for this is that, in this way, the performance of participant can be compared against more or less ‘advanced’ formulas. Consequently, three linear models with different number of predictors were established and were based on the following principles:

- A linear model that incorporated a number of attributes that yield the best result (see section 1.1)
- Two alternative linear models that include attributes that are considered as important contributors to a prediction on throughput time, by operators (see section 1.1)

2.4.2 Heuristics

Snijders et al. (2003) have shown that people working in a particular domain tend to use less information, and that they use information in a more combined non-linear way. For the lithographic industry, this suggests that people use different strategies to make a prediction.

Indeed, results from an explorative study brought up two interesting aspects. That is, the nominal print capacity as specified by the manufacturer is often used by managers whereas operators merely tend to make use of rule-of-thumbs. However, to restrict the number of different methods to be included, it was decided to focus only at the most common ones (i.e. mentioned by more than five participants).

2.5 Clinical versus statistical prediction

A comparison of the prediction on the total throughput time as predicted by the participants as well as those generated by the statistical formula will provide an answer on the research question. This section start by describing three criteria as measure of performance. Subsequently, the study design and the approach for the analysis of the data are described.

2.5.1 Performance indicators for a prediction

Since there is not a single measure that is ready to be used to determine how good or bad a prediction regarding throughput time is, three different criterions were defined. Each criterion is an objective measure of performance and allows for a comparison of both throughput times as predicted by the participants as well as the formulas. Moreover, considering different criteria related to a prediction of throughput times increases the external validity of the study and helps to strengthen claims later on. Note that for each criterion an average score is calculated from a set of eight vignettes.

2.5.1.1 Criterion 1: rank correlation between the actual and predicted throughput times

The *Spearman Rank c*orrelation is used as a measure to assess a participant's (or model) ability to rank a set of vignettes. Using the actual throughput times, as obtained from the printer data, a score is calculated that represents how well the ranking of the vignettes predicted by the participants (or model) matches those based on the actual throughput times. That is, a participant who ranks the vignette perfectly (from shortest to longest) would obtain a score of +1, whereas a participant that does this exactly the opposite obtains a score of -1 (closer to +1 is better).

2.5.1.2 Criterion 2: relative deviation between actual and predicted throughput times

The relative deviation is used as a measure to assess how accurate a participant (or model) predicts the total throughput time of a set of vignettes. The accuracy of a prediction is measured by calculating the difference between the predicted throughput time (by participant or the model) and the actual throughput time. Note that we do not distinguish between over / underestimation and therefore the absolute difference is taken. A skilled participant would obtain a score of 0 (closer to zero is better).

2.5.1.3 Criterion 3: number of vignettes predicted correctly

In contrast to the relative deviation, this criterion allows for an estimation of the prediction on the total throughput time for a set of vignettes. The range for which a prediction is counted as correct is set on a deviation of maximum 10 percent from the actual throughput time. The value for this range was obtained from the interviews that were conducted at the very start of this research. As it turned out, four operators thought that they were able to predict throughput time correct with a 10 percent deviation from the actual throughput time. A participant (or model) who predicts throughput time for all eight vignettes correct obtains a score of eight, whereas a participant who does not predict a single vignette correct obtains a score of 0 (closer to eight is better).

2.5.2 Study design

For the clinical condition, a vignette experiment was conducted for which participants were asked to make a prediction on throughput time for a set of print jobs. In addition, for the same vignettes a prediction on throughput time was generated by different formulas. Given that the actual throughput time of vignette is known from the data, every prediction is evaluated against three different criteria. A comparison of both the scores on the different criterion as obtained by the groups of participants *(clinical prediction)* and the ones as generated by the formula *(statistical prediction)*, will provide an answer on the research question.

2.5.3 Analysis

Two different software packages were used to analyze the data. That is, Microsoft Excel was used to prepare the raw printer data so that it could be used as input for the vignette experiment and to create a dataset that was ready to establish the statistical models upon. The statistical software package SPSS was used to calculate a three formulas that each generate a prediction on throughput time.

Moreover, SPSS was used for the comparison of the scores on the different criteria for both groups of participants and models. It should be noted that the scores on the first and second criterion were not normally distributed, so parametric testing is not allowed. Therefore, it was decided to compare groups and models non-parametrically. This method of testing has the advantages that it makes no assumptions about the type of data on which it can be used. Moreover, most of the test work under the principle of ranking the data, that is, the lowest scores in the data get a value of 1, then finding the next highest score and giving it a rank of 2, and so on. Eventually the analysis is carried out on the ranks rather than on the real data. It is claimed that this method has less statistical power than parametric testing. Therefore, finding a significant result makes it more likely that the effect is actually present when data was distributed normally (Field, 2005, p. 521). The following methods were used to compare the group and models statistically:

- *Mann Whitney U test*; is a non-parametric test for assessing whether two independent sample of observations have equally large values. This test will be used to compare the performance of the different groups of participants (Field, 2005).
- *Wilcoxon Signed Ranks test*; is a non-parametric test for the case of two related samples or repeated measurements on a single sample. This test will be used to compare the performance of participants against the different models (Field, 2005).

3. Printer data as input for this research

Since no data was readily available, a company needed to be found where a large number of documents were printed. In the upcoming paragraph, the characteristics of this company and a description of the print production process are presented. Subsequently, it is decribed how the data was checked on errors, processed and used for the most elementary part of this research, that is, to calculate a formula that generates a formula on throughput time *(statistical prediction)* and for the design of an experimental task *(clinical prediction)*.

3.1 Environment

Over a five-week period, data collection took place at the print room from a large insurance company in the Netherlands. The print load at this site, varying from 150.000 to 450.000 impressions per day, is allocated to three identical printers. More specifically, the printer that we collected data from is the Xerox Nuvera 288 (see Figure 1). According to the specifications, it has a nominal speed of 288 impressions per minute (IPM) for two-sided prints, whereas this rate drops to a 144 IPM for one-sided prints[1]. The jobs that were printed were relatively straightforward, for example invoices, policies and contracts, and was printed on A4 75 grams/m^2 (90%) and A4 90 grams/m^2 (10%) paper.

Figure 1 Xerox Nuvera 288 – From left to right the printer contains the following compartments: paper input (2x), control panel (top) and the two print engines (bottom), an extra paper input and an output (2x).

3.2 Data clean-up

For the purpose of data collection, the account log files of the printers were enabled. This functionality automatically keeps a record of every job that is printed. It contains information about start and stop times, document characteristics and more. However, before the data could be used it was checked manually on completeness and possible coding errors. The following procedure was carried out:

- The data file was opened in Microsoft Excel and only the parameters that contributed to a prediction on throughput time were kept (see Table 1).
- Next, some minor coding errors, due to paper jams, were removed. In addition, pauses between print jobs that could not be explained (e.g. mechanical failure, coffee breaks or whatsoever) we deleted whereas interruptions that appeared plausible were shortened to 25 minutes utmost (i.e. average time to prepare the printer).
- Lastly, to run a multiple regression analysis the data needed to be transformed so that a single row represented a complete day of work (instead of a detailed overview per day). Using the pivotal

[1] More information is found on: http://www.xerox.com/digital-printing/printers/print-on-demand/xerox-nuvera-288-digital-perfecting-system/enus.html

table functionality in Microsoft Excel, jobs with similar document names were grouped and the values of the attributes were totaled.

A closer look at the data showed that the total throughput time of a set of print jobs varied between 2 and 11 hours. The advantage of this is that it becomes less likely that a participant predicts the total throughput time of a vignette by means of guessing (e.g. as when every vignette represents a standard 8 hour working day). Consequently, the outcomes of one's predictions are a more objective measure of one's ability to predict throughput time.

Table 1 the parameters and its functionality used from the account log file. Parameters with an asterisk are visible for the operator when making a prediction on total throughput time.

Parameter	Functionality
Document name*	*The key identifier of a job. This parameter contains information about the client (and the type of paper that is needed). Print jobs that have similar document names are totaled.*
Number of impressions*	*Indicates how many impressions the job contains.*
one-sided sheets*	*Indicates how many sheets were printed 1–sided (1 impression).*
two-sided sheets*	*Indicates how many sheets were printed two-sided (2 impression).*
Medium type*	*Represents the different of paper that are required for the print job.*
Printer server name	*Contains the name of the printer the job was assigned too.*
Timestamp printing started	*Contains date and start time of when the job printed.*
Timestamp printing ended	*Contains date and end time of when the job was ready.*

3.3 Subdivision of the printer dataset

In total 75 days of data were collected in a five-week period (25 days x 3 printers). Approximately two-third of the dataset was set aside and used to calculate a formula that generates a prediction on throughput time. The remainder was used for the design of the vignettes *(clinical prediction)* and for validation of the formula *(statistical prediction).* That is:

1. 43 days were used to establish the statistical formulas using linear regression analysis.
2. 32 days were used to design vignettes for which a prediction on throughput time was made both by the participants and by formulas.

To be perfectly clear, note that the formulas as calculated from set 1 are used to generate a prediction on throughput time for the same vignettes as the ones that are used to assess the participants (i.e. set 2). In this way, we are able to compare the clinical and statistical conditions. Note that this kind of cross-validation to generate a formula that is to compete with human experts is common in more 'clinical versus statistical' studies (Snijders et al., 2003). In the subsection it is described how the statistical models were established and how the vignettes were designed.

3.3.1 Properties of the formulas

In total five models were incorporated in this study from which some details are found in Table 2 (a complete overview of the linear model is found appendix – A.) The first three models that are presented in this table were established using a multiple regressions analysis that. The fourth model is a heuristic that has been brought up by operators whereas the last model represents the nominal print capacity, as it specified by the manufacturer of the printer.

For the linear models, Table 2 provides an indication of the goodness of fit and the standard error. The goodness of fit represents how much of the variability in the data can be explained by the model. In addition, the standard error provides an indication of how much the predicted throughput time deviates from the actual time. For example, Table 2 shows that the accuracy increases proportionally as a function of the number of predictors. That is, the four-predictor linear model generates in 80 percent of the cases (vignettes) a throughput time that deviates 61 minutes utmost from the actual time whereas for the remaining 20 percent the model cannot provide a reliable indication. However, the fact that the best model captures (only) 80 percent of the variability that is in the data implies that there are some predictors missing. The reason for this is because, usually, mechanical processes such as printing a relatively predictable and a goodness of fit of 90 to 95 percent is therefore more likely. Interviews that were conducted as part of the explorative study, pointed out that the remaining variability can be attributed to interpersonal differences among operators who worked at the site were data collection took place.

Alternatively, Table 2 shows that properties of the heuristic. That is, the actual print capacity is estimated on 10.000 impressions per hour instead of a nominal value of 17.280. This provides a good example of how is account for downtime in practice. Lastly, Table 2 shows the properties of the nominal model. The reason for including this method is that it was often mentioned by students and some managers. Note that for these two alternative methods the average deviation from the actual throughput time and the standard deviation are also shown.

Table 2 Properties of the different models. In contrast to the statistical models, we can provide the average deviation from the actual time and the standard deviation for the last two models.

Model	Predictor	Type of model	Goodness of fit	Standard error
four-predictors	*Jobs, one-sided sheets, two-sided sheets, paper trays*	linear regression	80%	61 minutes
1-predictor (imp.)	*Total # impressions*	linear regression	63%	78 minutes
1-predictor (two-sided)	*Total #two-sided sheets*	linear regression	71%	91 minutes
	Predictor	**Type of model**	**Average Deviation**	**Standard Deviation**
Heuristic	*10.000 impressions per hour*	‘rule-of-thumb'	60,7 minutes	39,3 minutes
Nominal	*8640 sheets per hour*	nominal model	118,9 minutes	60 minutes

3.3.2 Construction of the vignette experiment

The goal of the vignette experiment was to test participants on their ability to make a prediction on throughput time for a set of print jobs. Hence, the experimental task was made up of the following sections:

- *Demographics* – which were used to get more complete understanding of the obtained results, see appendix – B.
- *Task description* – which included to explain the purpose of the study and to provide a background of the print process the data originated from, see appendix – C.
- *Vignette experiment* – which contained a set of eight vignettes for which throughput time needed to be predicted. In total 32 vignettes were created (see section 3.3). Vignettes were randomly divided into four sets of eight vignettes. Every set was assessed by eight participants per group (managers, operators, and students). Moreover, order and sequential effects were mitigated because every participant received a set in which vignettes were ordered randomly. After a set was assessed by eight participants per group, it was discarded from analysis.
- *Method section* - where participants were asked to write down the method they were using, see appendix – D.

3.3.2.1 Measures

In order to obtain a reliable measure of one's level of expertise, Ericsson (2003) proposes to restrict the scientific evidence to those aspects that can be repeatedly and reliably observed, such as the variability in the outcome. Accordingly, per vignette two representative measures were included. The first measure captures an individual's ability to predict throughput time for a set of print jobs whereas the second measure an individual's ability to reflect on their own performance.

- Participants had to predict the total throughput time for a vignette (see Figure 2). Recap that throughput time was defined by the total time it takes to print a number of jobs *(machine time)* plus some amount of time for when the printers status is idle *(down time)*.
- Participants had to indicate how much they thought their prediction deviates from the actual throughput time that was known from the print data (see Figure 2). Note that this was measured on a six-step categorical scale, ranging from 10 minutes to more than 1.5 hour.

3.3.2.2 Design

The vignette design had to represent a self-contained overview that closely matched an ordinary work planning. The layout of the task for a single set of jobs is shown in Figure 2. This overview offers a slight advantage to an ordinary work overview, because we totaled the values for the most important job characteristics already. Moreover, the first column shows the number of sets (i.e. different clients) and job ID. The second column shows the total number ofjobs that need to be printed. The third column shows the total number of impressions. The fourth and fifth column show the number of one-sided and two-sided sheets. Finally, the last column shows the different types of paper that are needed for a particular job.

A pilot study with three operators and three students was conducted, in order to check whether the task was understandable and easy to administer. Based on these findings, we concluded that the purpose of the task was clear, did not take longer than 15 minutes to complete and that it was ready for the full-blown study.

49

Set 1	**Totalen Opdrachten**	**Afdrukken**	**Enkelzijdige bedrukte vellen**	**Dubbelzijdige bedrukte vellen**	**Te gebruiken papierladen**
polispapier	**73**	**72994**	**1650**	**35672**	**4**
430X102	28	7185	459	3363	4
430X97D	15	22924	202	11361	3
430X98D	30	42885	989	20948	3
briefpapier	**2**	**98**	**34**	**32**	**2**
430X085	2	98	34	32	2
Totalen	**75**	**73092**	**1684**	**35704**	**4**

Geschatte benodigde tijdsduur voor alle opdrachten samen (omsteltijd + machinetijd):
(geef een zo precies mogelijke schatting)

.... uur en .... minuten

Hoe goed verwacht u dat uw schatting is (omcirkelen wat van toepassing is):

o Hooguit 10 minuten ernaast
o Hooguit 30 minuten ernaast
o Hooguit 45 minuten ernaast
o Hooguit 1 uur ernaast
o Hooguit 1.5 uur ernaast
o Mogelijk meer dan 1.5 uur ernaast

Figure 2 a vignette of a print job for which participants were asked to predict the total throughput time need.

4. Results

This chapter describes the results from the vignette experiment and compares them against the scores from the different formulas. First, details about the participants are provided. Next, an overview of the performance for the different groups *(clinical prediction)* and models is presented *(statistical prediction)*. From here on, the results from a comparison between the performance of the different groups and formulas is described. Note that every result section ends with intermediate conclusions and implications. In this section, we briefly elaborate on the most prevalent findings and new questions will be raised that are addressed by a more thorough analysis of the data.

4.1 Participants

Difficulties were encountered with finding companies that were willing to participate in this study. Nevertheless, 71 participants were tested. The group that was strongly underrepresented were managers because there only were a few of them working at each company that was visited. In addition, the data from two operators and one manager had to be discarded from analysis since they predicted similar throughput times for every vignette. That is, the manager and an operator predicted a throughputs time of almost 10 hours for every vignette whereas the other operator predicted an hour for every vignette. Moreover, it turned out that finding operators that had at least ten years of work experience was also more difficult than expected. Therefore, it was necessary to lower this criterion. Consequently, an average of 10 years of work experience (without a lower limit), was accepted as suitable. The characteristics of the remaining participants were as follow:

- *Students:* 27 participants, 24 were male and three were female (M=23 years, SD= 7.6)
- *Operators:* 28 participants, 25 were male and three were female (M= 37 years, SD= 9.8).
- *Managers:* 13 participants, 8 were male and four were female (M= 42 years, SD= 4.9).

4.2 Descriptive statistics

To get a first impression of how the participants performed compared to the different models, the average scores on the different criterion were calculated per group. Table 3 shows that operators and managers perform quite similar, whereas students perform worse. In addition, note that the heuristic performed reasonable compared to the statistical models. Lastly, two other aspects that are worth mentioning here:

- ‘Perfect’ shows the score that would have been obtained in case the throughput time of the vignettes is predicted similar to the actual throughput time.
- The nominal model shows in what direction the scores on the different criteria go when not is accounted for downtime.

Table 3 Score on different criteria for different groups and models (model scores are averaged among groups). From top to bottom: perfect represents the optimal score that can be obtained, the middle section shows the scores for different groups, and the lower section shows the scores of the formulas.

Groups	Rank correlation	Relative deviation	Vignettes correct
'Perfect'	*1*	*0*	*8*
Students	*0,63*	*0,32*	*1,48*
Operators	*0,60*	*0,24*	*3,18*
Managers	*0,75*	*0,28*	*3*
4-predictors	*0,88 (σ = .05)*	*0,13 (σ ≤.01)*	*5,51 (σ = .02)*
1-predictor (imp.)	*0,63 (σ = .02)*	*0,21(σ ≤.01)*	*2,75 (σ = 0)*
1-predictor (two-sided)	*0,82 (σ = .06)*	*0,14(σ ≤.01)*	*4,51 (σ = .02)*
Heuristic	*0,82 (σ = .06)*	*0,17(σ ≤.01)*	*3 (σ = 0)*
Nominal	*0,79 (σ = .07)*	*0,33(σ ≤.01)*	*0 (σ = 0)*

4.3 Clinical prediction

In this paragraph, the results from the vignette experiment are presented. Not that for the groups that are significantly different, the effect size is denoted by the *Pearson's correlation* for non-parametric tests (Field, 2005). This is simply a measure an objective and standardized measure of the magnitude of the observed effect. The following values are in general accepted about constitutes a large of small effect:

- r = .10 (small effect): in this case, the effect accounts for 1% of the total variance.
- r = .30 (medium effect): the effect accounts for 9% of the total variance.
- r = .50 (large effect): the effect accounts for 25% of the total variance.

4.3.1 Students versus professionals

For the first analysis, students are compared against professionals (i.e. managers and operators). In line with our expectations, it was found that students had a larger relative deviation and predicted fewer vignettes correct, whereas they scored indifferent on the rank correlation. That is, The *Mann-Whitney U test* showed a distribution of scores on the different criteria between students and professionals is as follow:

- **Rank correlation:** students (M= ,71) versus professionals (M= ,73)
  U= 525; p=,789
- **Relative deviation:** students (M= ,33) versus professionals (M= ,27)
  U= 358,5; p=,017; r= -, 11
- **Number of vignettes correct:** students (M= 1) versus professionals (M= 3)
  U= 297; p=,001; r= , 15

With one exception, these results show that professionals (i.e. operators and managers) perform significantly better than students do. Note that the effects size is rather small.

4.3.2 Operators versus managers

Given the fundamental differences in work activities (feedback versus no feedback); one would expect that mangers are outperformed by operators. Results from the *Mann-Whitney U test* show that the distribution of scores on the different criteria between operators and managers is as follow:

- **Rank correlation:** operators (M= ,70) versus managers (M = ,73)
  U= 115,5; p=,157
- **Relative deviation:** operators (M= ,22) versus managers (M= ,27)
  U= 148; p=,670
- **Number of vignettes correct:** operators (M= 3) versus managers (M= 2)
  U= 158; p=,902

This result shows that operators perform similar to managers and implies that frequent and direct feedback has no effect.

4.3.3 Students versus managers

In section 4.3.1, it was shown that professionals outperform students. However, it can still be that this effect largely is attributed to operators because the two groups differ considerably in the number of participants (i.e. 28 versus 13). To rule out this possibility the *Mann-Whitney U test* shows that the distribution of the scores on the different criteria between students and managers is as follow:

- **Rank correlation:** students (M= ,71) versus managers (M= ,73)
  U= 119; p = ,274
- **Relative deviation:** students (M= ,33) versus managers (M= ,27)
  U= 118,5; p= ,144
- **Number of vignettes correct:** students (M= 1) versus managers (M= 2)
  U= 91; p= ,022; r=, 15.

This result show that managers outperform students only on the total number of vignettes correct, whereas they performed equal on the others tasks. Furthermore, this finding suggests that the effect, as it was found in the first contrast, is largely attributed to operators' performance.

4.4 Comparison between clinical versus statistical predictions

In this section, the results are described of a comparison between participants and the different formulas. In line with our research question, cases in which the distribution of the data for operators is significantly better than those of the formulas are reported. In addition, cases for which a model performs similar to a group of participants are reported without significance level.

Given that the data was not normally distributed, the *Wilcoxon Singed Ranks test* was used to check for significant differences in the distribution of scores. For every criterion, the operators' performance was compared against the various models and was repeated for the managers. For criterion where managers and operators performed indifferent, as known from the results in section 4.3, the group scores were taken together. Note that a complete overview of the test statistics is found in appendix – E.

4.4.1 Rank correlation

In section 4.3.1, it was concluded that students scored similar to professionals (i.e. operators and managers) on the first criterion. Therefore, in this analysis the scores for these groups were taken together and compared against the different models (Figure 3). The *Wilcoxon Signed Rank test* (Z= -, 996; p=, 319) shows that the distribution of the scores for participants (M=, 71) equals that of a linear model with impressions as a predictor (M=, 69), whereas the distribution of the data for all other models was significantly better on the, 05 level. This finding shows that humans are not able to outperform a formula.

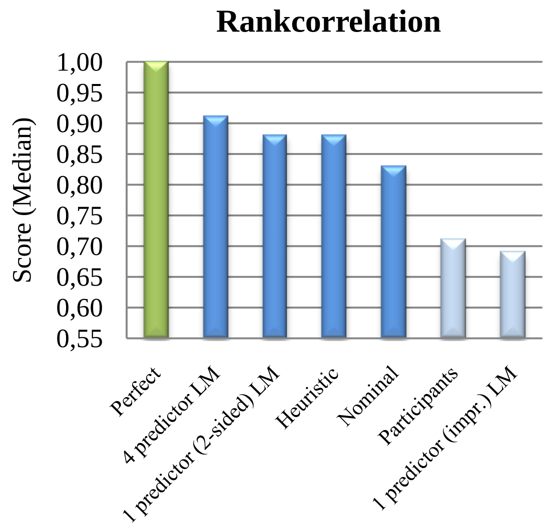

Figure 3 Median scores for groups and models (higher = better). Note that similar colors represent groups (or models) that cannot be distinguished from each other statistically.

4.4.2 Relative deviation

In addition, in section 4.3.1 it was concluded that students performed different from professionals on the second and third criterion. Therefore, for this analysis these groups were analyzed separately (Figure 4). The *Wilcoxon Signed Ranked test* showed a non-significant result (Z= -, 751, p=, 452) in the distribution of scores for students (M=, 33) and that of the nominal model (M=, 34), indicating that they perform similar. On the contrary, the distribution of scores for professionals (M=, 26) was significantly better (Z= -3,770, p=, 006) than the nominal model (M=, 33). Moreover, a marginally significant effect (Z= -1,891, p=, 059) was found between professionals (M=, 26) and the linear model with two-sided sheets as a predictor (M=, 20), indicating that they perform similar. Lastly, the distributions of the data for all other models were significantly better on the .05 level. Taken together, these results show that professionals perform better than the nominal model and similar to a one-predictor linear model with two-sided prints as predictor, whereas they are outperformed by all other models.

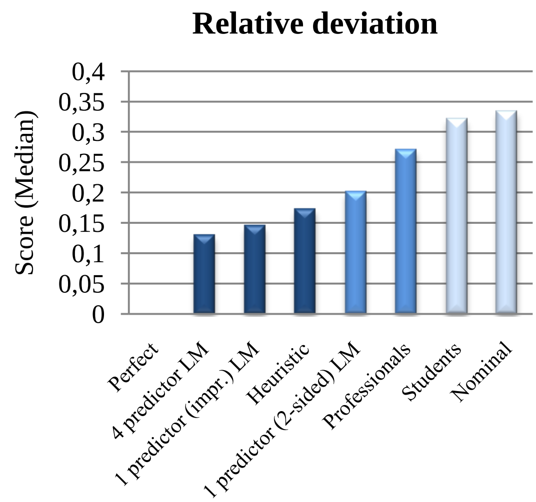

Figure 4 Median scores for students and professionals versus models (lower = better). Note that similar colors represent groups (or models) that cannot be distinguished from each other statistically.

4.4.3 Number of vignettes correct

As shown by the results from the clinical prediction, managers performed indifferent from operators and therefore these groups are taken together. However, for the third criteria the results are somewhat different compared to what has been found so far. That is, *the Wilcoxon Signed Ranks test* showed that the distribution of the data was significantly better (Z= 3, 97; p=, 00) for students (M= 1) compared to the nominal model (M= 0). Likewise, the distribution of the data was significantly better (Z= -5,184; p=, 000) for professionals (M= 3) compared to the nominal model (M = 1). Furthermore, professionals (M= 3) performed non-significantly different (Z= -1,599; p =, 110) from the linear model with two-sided sheets as a predictor (M= 3) and performed non-significantly different (Z= -, 167; p =, 867) from the heuristic (M= 3). Lastly, the distributions of the data for all other models were significantly better on the .05 level (see Figure 5). In conclusion, these results show that professional perform similar to a heuristic and a one-predictor linear model with the number of two-sided sheets as predictor, whereas they are outperformed by two other models.

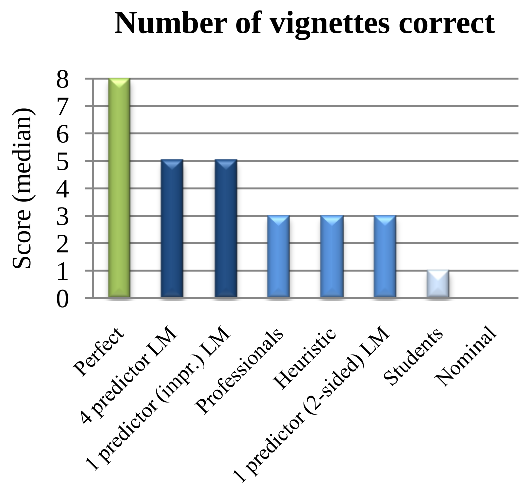

Figure 5 Medians scores for students and professionals versus models (higher = better). Note that similar colors represent groups (or models) that cannot be distinguished from each other statistically.

4.5 Intermediate conclusion

The findings so far yield to the same conclusion that is: [*1] professionals do better than students; [2] managers perform similar to operators and [3] professionals are outperformed by relatively simple models, including a commonly used heuristic (only on the second and third criterion).* In the upcoming section, we briefly elaborate on these conclusions and propose new questions for a more complete understanding of the obtained results.

4.5.1 Implications for further analysis

First, the fact that professionals perform better than students is something that was expected beforehand. The reason for this is that students are not familiar with printing and therefore rely on the information that was provided to them. However, Figure 4 and Figure 5 in the last section suggest that students did not solely use the nominal print capacity but accounted for downtime to some extent. Consequently, they performed slightly better compared to the nominal model but still worse than the professionals did.

Secondly, the fact that managers perform similar to operators is perhaps most striking with our expectation. Due to the fundamental different working activities (feedback versus no feedback), it was expected that operators would outperform managers, yet this appeared not to be true. A possible explanation for this finding comes from personal observations during site visits. That is, managers frequently inquire at the printing department when dealing with planning issues. In this way they get to know the relevant aspects of the process and learn how make a prediction on throughput time, without being confronted with the outcomes. Taking this one-step further, it suggests that managers' performance does not depend on work experience whereas that from operators does. Likewise, Snijders et al. 2003 found that older (senior) managers were not able to outperform their younger colleagues and attributed this finding to a lack of frequent and direct feedback. In addition, the literature has learned us that frequent and direct feedback contributes to a better performance (Ericsson, Krampe, & Tesch-Romer, 1993).Taken together, this implies that frequent and direct feedback mediates the relationship between work experience and an operator's ability to make a prediction on throughput time. Hence, we formulate the following sub research question:

> *[RQ 1.1] Does frequent and direct feedback allow more experienced operators to outperform their younger colleagues, when predicting throughput time for a set of print jobs?*

Furthermore, Early, Northcraft, Lee, & Lituchy, (1990) state that outcome feedback allows individuals to reflect upon their behavior with predefined goals and to determine whether to adjust their actions. More specifically, people use discrepancies between goals and outcome feedback as the basis for self-reflection that eventually leads to better performance. This implies that frequent and direct feedback mediates the relationship between self-confidence and an operator's ability to make a prediction on throughput time. Hence, we formulate the following sub research question:

> *[RQ 1.2] Does frequent and direct feedback allow more self-confidence operators to outperform their less self-confident colleagues, when predicting throughput time for a set of print jobs?*

So far, it is concluded that frequent and direct feedback does not allow operators to outperform a statistical model. However, before we reject the main research question, it was aimed to rule out one last possibility. That is, if we would find that more experienced or self-confident operators perform better than their colleagues do, than it is of particular interest to test whether they are able to outperform a statistical formula. Only if it appears that neither the most experienced nor the most confident operators are able to outperform a statistical formula, than we have provided ample arguments to conclude in favor of the statistical method.

4.6 Supplementary analysis of the data

The sub research questions, as formulated in section 4.5.1, require a more thorough understanding of the data. The following methods were used for this:

- *A bivariate correlation (Spearman Rho)* – used to measures the strength of a relationship between the two variables. The variables that were obtained from the demographics was included as part of the vignette experiment. The strength of a correlation ranges from zero to one; the closer the value is to one, the stronger the relationship; the closer and the value is to zero, the weaker the relationship. In addition, a positive or negative sign provides an indication in what direction the relationship goes.
- *A multiple regression analysis* – used for testing whether the bivariate relationship remains to exist or if the effect is mitigated when other predictors are included in a linear model.
- *Wilcoxon Signed Ranks test* – used for a comparison between the performance of participants and the different models.

4.6.1 The effect of work experience on performance

Given that operators are used to frequent and direct feedback on the outcomes of their predictions, the more experienced operators should make a reasonable chance to outperform a statistical formula. Subsequently, in the subsections it is tested whether this proposition holds or not.

4.6.1.1 Bivariate correlation between work experience and performance

In order to test whether a relationship between work experience the different criteria exist, the S*pearman Rank* correlation is calculated. The strength of the relationship is shown by the Rho-value. The results are as follows:

- *Rank correlation:* a non-significant relationship was found between the rank correlation and work experience for both operators and managers.
- *Relative deviation:* only for operators a significant relationship was found between the relative deviation from the actual time and years in printing (Rho =, 43; p =, 025).
- *Number of vignettes predicted correctly:* only for operators, a significant relationship was found between the number of vignettes correct and years in printing (Rho =, 399; p =, 039).

Taken together, these results suggest that frequent and direct feedback allows more experienced operators to outperform their younger colleagues. However, an additional analysis is required in order to strengthen this claim.

4.6.1.2 Multiple regression analysis on work experience

A *multiple regression analysis* was conducted in order to test the robustness of the significant correlations. To our surprise, a an initial analysis with as dependent variable the *relative deviation from the actual throughput time* (criterion 2) or *number of vignettes correct* (criterion 3) and work experience as a predictor, did not yield a significant result. This problem was solved after removing the data from one of the most experienced operators, who apparently predicted throughput times much too high.

Next, all demographics that were obtained were included as predictors in a linear model, upon which a multiple regressions analysis was conducted. However, some predictors had to be discarded from the model because they correlated highly with work experience (see Table 4). This phenomenon, also referred to as multi-collinearity, occurs when two of more variables measure essentially the same thing (Field, 2005). The problem is normally solved by finding a way to combine variables or by removing the ones that do not contribute logically to the model. Due the small number of operators (N=28), the only remedy in this study was to remove the variables and understand how they possibly affect the independent variable.

Possible explanations for why some variables correlated highly with work experience can be provided by looking at Table 4. First, age and work experience coincide and therefore it is obviously why they correlate highly. Secondly, work experience and profession correlate highly because, typically, an operator starts his career on a low level job and slowly climbs up to become a key-operator. Likewise, due to a possible effect of frequent and direct feedback, an operator is able to reflect upon their performance, as a result operators start to feel more confident and makes better predictions as years in printing increases.

Table 4 Predictors that were obtained from the demographics and its Pearson correlation with work experience. Note that the ones denoted with an asterisk were discarded from the model, because they correlated highly with work experience (i.e. multicollinearity).

Predictor	Definition	Work experience
Hours per week	*Hours per week working with the printer*	P = ,096
Years	*Years working with the printer*	P = ,116
Finishing applied	*Printer used for finishing of the document*	P = -,110
Gender	*Of the operator / manager*	P = ,370
Counter	*Operator / manager responsible for registration of total impressions*	P = -,028
Age	*Age of the operator / managers*	P = ,705*
Profession	*(Key) Operator / all-round operators / manager*	P = ,717*
Self-confidence level	*Measure for making prediction (pre-, and post of the experiment)*	P = -,378*
Average deviation	*Estimate of how far a prediction deviates from the actual throughput time*	P = -,374*

After the variables that caused multicollinearity were removed from the model, a multiple regression analysis on work experience was conducted. Note that, a non-significant bivariate relationship between the rank correlation and work experience was found in section 4.6.1.1, and therefore this analysis was only focused on the second and third criteria. The results are as follows:

- *Relative deviation:* a linear regression analysis with relative deviation from the actual throughput time (criterion 2) as the dependent variable and the predictor that remained (see Table 4), resulted in a significant model (F (6, 24) = 2,784; p=, 045), with work experience as a marginally significant predictor (standardized-β=, 404; p=, 057).
- *Number of vignettes predicted correctly:* a linear regression analysis with the number of vignettes correct (criterion 3) as the dependent variable and the predictors that remained (see Table 4), resulted in a not-significant model (F (6, 24) = 2,338; p=, 076), yet with work experience as a significant predictor (standardized-β =, 624; p=, 007). Note that the non-significance indicates that there is an imperfect fit between the model and the data. Consequently, the R-square (amount of variance explained by the model, also referred to as goodness of fit) might be biased and needs to be interpreted carefully (Field, 2005).

The most relevant outcome of a regression analysis is that it gains insight into the ‘importance’ of a particular predictor in the model (Field, 2005). The best way to interpret this is by looking at the standardized-β values, which represents the strength the relationship. Given that these values are all measured in standard deviation units, they are directly comparable.

With an exception for rank correlation, we conclude that their, indeed, is a positive relationship between work experience and performance. This suggests that more years in printing lead to a more accurate predictions on the total throughput time for a set of print jobs. As we do not find this result for managers, we attribute this to an effect of frequent and direct feedback. In accordance with the literature this finding reconfirms a general conclusion that adequate feedback is essential in order to improve performance (Ericsson et al., 1993; Snijders et al., 2003) However, for this research it is of particular interest whether this selected group of operators is able to outperform a statistical formula. Subsequently, the next section provides the results of how they performed compared to the formulas.

4.6.1.3 Experienced operators versus models

In this section, the results are described from a comparison between the most experienced operators and the different formulas. Operators (N=8) with at least 10 years of work experience were deliberately selected and for the comparison the non-parametric paired Wilcoxon Singed Ranks test was used.

Relative deviation: for the second criterion (see Figure 6) it was found that the distribution of the data for operators was significantly better (Z = -2,521; p= ,012) than the nominal model (M= ,35). On the contrary, the distribution of the data was significantly worse (Z= -2,1; p =,036) for operators (M= ,18) and the four-predictor linear model (M= ,12). Furthermore, the distribution of the data for operators was non-significantly different from the models that remain, indicating that they perform similar. Hence, this finding suggests that more experienced operators perform similar to three models but that a four-predictor linear model outperforms them.

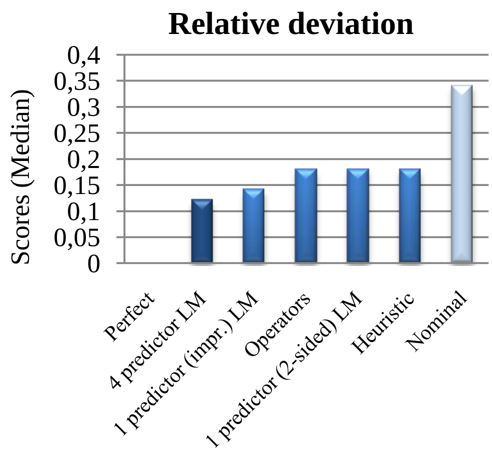

Figure 6 median scores for operators with at least 10 years of work experience versus models. Note that similar colors represent groups (or models) that cannot be distinguished from each other statistically.

Number of vignettes correct: for the third criterion (see Figure 7) the distribution of the data showed that operators perform significantly better (Z= -2,539; p= ,011) than the nominal model (M= 0). Likewise, the distribution of the data for operators was significantly better (Z= -1,98; p= ,048) than the heuristic (M= 2). Lastly, the distribution of the data showed that operators perform significantly better (Z= -2,271; p= ,023) than the linear model with two-sided sheets as predictor (M= 3). On the contrary, the distribution of data showed that operators (M= 5) perform significantly worse (Z= 1,983; p= ,047) than for the four-predictor linear model (M= 6). Furthermore, the distribution of the data for operators was non-significantly different from the one predictor model with impressions as a predictor, indicating that they perform similar. Taken together, these results show that the more experienced operators outperform three models and perform similar to another, but that a four-predictor linear model outperforms them.

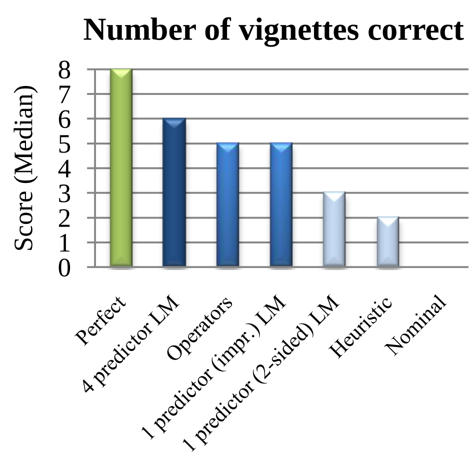

Figure 7 median scores for operators with at least 10 years of work experience versus models. Note that similar colors represent groups (or models) that cannot be distinguished from each other statistically.

4.6.2 The effect of self-confidence on performance

Given that operators are used to get frequent and direct feedback on the outcomes of their predictions, a sub research question was formulated in section 4.5.1. Following the same procedure as previously, it is tested whether there is a relationship between self-confidence and performance. Subsequently, the robustness of this relationship is tested using a multiple regression analysis. Lastly, the operators' performance is compared against that those of the different formulas.

4.6.2.1 Bivariate correlation between self-confidence level and performance

Prior and post to the vignette experiment participant were asked how confident they were about their own predictions. The *Spearman Rank* correlation was calculated to test whether a relationship exists the average score was calculated for self-confidence level and correlated with the different criteria. The results are as follow:

- *Rank correlation:* a non-significant relationship was found between the rank correlation and self-confidence level for both operators and managers.
- *Relative deviation:* only for operators a significant relationship was found between the relative deviation and self-confidence level (Rho =, 422; p=, 036).
- *Total number of vignettes correct:* only for operators a (marginal) significant relationship was found between the number of vignettes correct and self-confidence level (Rho =, 387; p=, 056).

Taken together, this finding suggests that accuracy of the prediction increases proportionally as a function of confidence level for operators, yet not for managers. However, an additional analysis is required in order to strengthen this claim.

4.6.2.2 Multiple regression analysis on self-confidence level

In order to test whether the bivariate relationship is robust, a multiple regression on self-confidence level was conducted. For the same reason as stated earlier, it was found that multicollinearity was also present here (details are found in section 4.6.1.2 - Table 4). Note that, a non-significant bivariate relationship between the rank correlation and self-confidence level was found in section 4.6.1.1, and therefore this analysis was only focused on the second and third criteria. The results are as follows:

- *Relative deviation:* a linear regression analysis with the relative deviation from the actual throughput time (criterion 2) as the dependent variable, resulted in a not-significant model (F (5, 24) = 1,994; p =, 126), yet with self-confidence level as a significant predictor (standardized-β =, 492; p=, 022).
- *Total number of vignettes correct:* a linear regression analysis with the number of vignette predicted correctly (criterion 3) as the dependent variable, resulted in a non-significant model (F (5, 24) = 1,377; p =, 277), yet with self-confidence level as a significant predictor (standardized-β =, 467; p =, 039).

Please note that the non-significance of the models indicate that there is an imperfect fit between the model and the data and that therefore the R-square (goodness of fit, which is a measure of the amount of variance that is explained by the model) might be biased (Field, 2005).

Looking at the standardized-β values, it turns out that there is a positive relationship between self-confidence and performance. As this is not found for managers, this is attributed to an effect of frequent and direct feedback. In accordance with the literature, this reconfirms a general conclusion that feedback allows individuals to reflect upon their behavior and consequently improve performance (Early, Northcraft, Lee, & Lituchy, 1990). More interestingly of this research is whether this group of operators is able to outperform a statistical formula.

4.6.2.3 Self-confident operators versus models

In this section the results are presented from a comparison between operators who were highly confident in their predictions to those generated by different formulas. Using the non-parametric paired *Wilcoxon Singed Ranks test* operators who stated that they were quite confident (n=12) and definitely confident (n=6) were tested.

Relative deviation: for the second criterion (see Figure 8), the distribution of the data was significantly better (Z= -2,940; p=, 003) for operators (M=, 20) than the nominal model (M=, 33). On the contrary, the distribution of the data was significantly worse (Z=-1,916; p=, 055) for operators than the four-predictor linear model (M=, 13). In addition, the distribution of data different was also significantly worse (Z= -2,765; p=, 006) for operators and the linear model with impressions as a predictor (M=, 13). Lastly, the distribution of the data for the models that remain was non-significantly different on the .05 level. This suggests that operators' performance is comparable to them. In conclusion, these results show that two statistical formulas outperform self-confident operators.

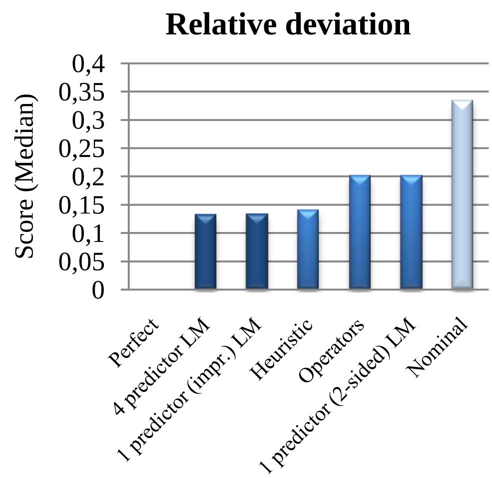

Figure 8 median scores for operators that were highly confident versus different models. Note that similar colors represent groups (or models) that cannot be distinguished from each other statistically.

Total number of vignettes correct: for the third and last criterion (see Figure 9), the distribution of data was significantly better (Z= -3,636; p≤, 01) for operators and the nominal model (M= 0). On the contrary, the distribution of data was significantly better (Z= -3, 24; p=, 001) for operators (M= 3, 5) than for the four-predictor linear model (M= 5). In addition, the distribution of data was significantly worse (Z= -2,456; p=, 014) for operators and the linear model with impressions as predictor (M= 5). Lastly, the distribution of the data for the models that remain was non-significantly different on the .05 level. In conclusion, this result shows that two statistical models outperform self-confident operators.

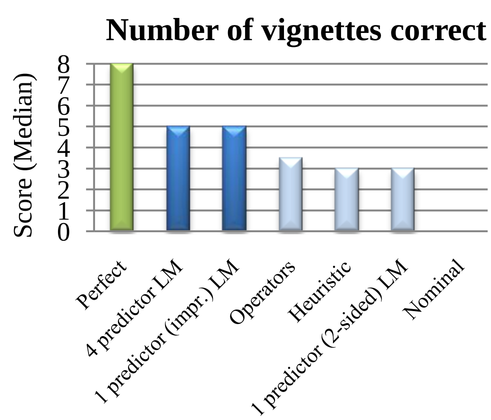

Figure 9 median scores for operators that were highly confident versus different models. Note that similar colors represent groups (or models) that cannot be distinguished from each other statistically.

4.7 Intermediate conclusions

The findings so far yield to the same conclusion that is: *[1] frequent and direct feedback allows more experienced operators to outperform their younger colleagues; [2] frequent and direct feedback allows operators to reflect upon their performance and to improve this; [3] neither the most experienced nor the most confident operators are able to outperform a four-predictor linear model.*

4.7.1 Implications for further analysis

Although the general conclusion so far is yielded in favor of the statistical approach, we want to rule out one last possibility before arriving at the overall conclusion. That is, it still can be that there are situations in which operators' expertise comes better to its right and consequently they are able to outperform a statistical formula. Hence, we can imagine that operators' predictions are better than statistical models for extraordinary situations. In the last section of this chapter we investigate if this proposition holds or not.

4.8 Robustness of the model

The last section ended with the proposition that performance of the models may be affected by changes in the characteristics of a set of print jobs. Hence, two aspects that might have an effect on the total throughput time of a set of print jobs are: *[1] when print capacity decreases* or when *[2] downtime increases.* The following arguments can be made for this:

- *Print capacity* is influenced by an increase in the number of one-sided impressions. Since the printer's capacity is expressed in impressions per minute (IPM), this value drops by 50 percent for one one-sided sheets. The reason for this is that the print capacity for one-sided prints is half that of two-sided prints (288 versus 144 IPM).
- *Downtime* is indirectly influenced by the total number of paper trays that are in a set. As this number increases, paper needs to be changed more often. Consequently, the chance that the printer status is idle becomes larger.

4.8.1 Descriptive statistics

In total, there were 32 vignettes, divided over four sets, for which throughput time needed to be predicted. The values of the most important characteristics per set were averaged and are shown in Table 5. Note that, for set 2 the ratio one-sided versus two-sided sheets *(print capacity)* as well as the largest number of paper trays *(downtime)* is highest among all four sets and also that the total number of impressions is relatively low compared to the rest. Hence, this suggests that if the proposition holds then this it would be likely that the performance of the models degrades for set two.

Table 5 Average values, calculated per set, for the characteristics of a set of print jobs.

Set	Ratio (one-sided / two-sided)	Paper trays	Jobs	Impressions	Total throughput time (min)
1	,43	15,75	116,5	64.453	414
2	,71	18,875	96	51.073	333
3	,26	16,75	84,5	58.584	355
4	,35	13,875	101,75	59.297	330

4.8.2 Operators versus models

As hypothesized, it was found that the performance of the statistical models is worse for set 2. However, this appeared to be true only for the second criterion (see Figure 10). That is, the non-parametric paired *Wilcoxon Singed Ranks test* showed that the distribution of the data for operators was significantly better (Z= -2,023; p=, 043) than the nominal model (M=, 3334) and the distribution of the data was (marginal) significantly better (Z = -1, 89; p=, 059) for the linear model with two-sided sheets as a predictor (M=, 28). On the contrary, the distribution of the data was significantly worse (Z= - 2,023; p =, 043) for operators (M=, 1930) than for the heuristic (M=, 1386). Lastly, the distribution of the data for the models that remain was non-significantly different on the .05 level. In conclusion, it shows that operators can reach the performance of a four-predictor linear model but that they are outperformed by a heuristic.

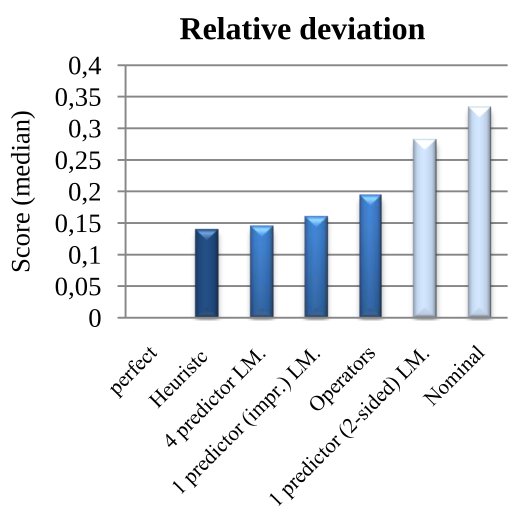

Figure 10 median scores for operators and the different models. Note that similar colors represent groups (or models) that cannot be distinguished from each other statistically.

4.9 Intermediate conclusion and implications

The results so far yield to the conclusion that is: *[1] operators reach performance of a four-predictor linear model but are not able to surpass it; [2] a four-predictor linear model is outperformed by a heuristic, however only under highly specific conditions.*

The fact that operators come as close to the performance of a four-predictor linear model suggests that the accuracy of the formula deteriorates when print capacity decreases. The magnitude of this effect becomes obvious when looking at the total number of impressions, in respect of the total throughput time per set. In contrast to set 4, we see that there are considerably more impressions printed while the total throughput time is more or less similar to that of set 2. This confirms that print jobs containing a relatively large number of 1 sided sheets and / or a large number of paper trays seem to affect the accuracy of the model and that operators are better able to anticipate on this.

In addition, the fact that the heuristic outperforms a four-predictor linear model, under highly specific circumstances, highlights the accuracy of this method. However, it also provides a good example why it is claimed that these rules are often inaccurate and easily misused (Shanteau, 1992, p. 192).

5. Conclusion and Discussion

This last chapter provides a summary of the overall research. First, the objective and main research question is reiterated, followed by the main points of the literature review. Subsequently, the scientific method that was carried out and the results are described. To that extent, a discussion is laid down in which is the pros and cons of this study are described followed by the scientific and practical implications. Lastly, some general shortcomings of this study are presented and directions for future research are given.

The objective of this research was to investigate whether individuals, who are used to get frequent and direct feedback on the outcome on their predictions, are able to outperform a statistical formula. The general consensus in the field of 'clinical versus statistical' prediction, associated with this type of research, is that humans are typically not good at making accurate predictions and often are outperformed by simple statistical formulas (Grove et al., 2000). However, a commonly stated objection to this conclusion is that in most studies feedback was lacking. Without adequate feedback on the outcomes of one's predictions, it becomes impossible for an individual to learn from one's own mistakes.

A literature review pointed out that feedback, indeed, play an important role in reaching optimal performance to make predictions in various domains (Ericsson et al., 1993; Graham, 1971; Hackman, Oldham, Janson, & Purdy, 1975; Snijders et al., 2003; Wright et al., 1996). In addition, Snijders et al. (2003) describes three common characteristics of environments in which statistical models seem to perform well. An explorative study was conducted to combine these findings into an environment that accounts for most of the aspects mentioned in the literature. As it turned out, the lithographic industry was found to be suitable for this research. Part of printing machine operators' work, involves making predictions on the time needed for a large number of print jobs, also referred to as total throughput time. Feedback is available in the form of a discrepancy between the predicted and the actual throughput time, and results sometimes in overtime work or a warning from the higher management. Hence, the following research question was formulated:

> *[RQ 1] Do humans outperform statistical models when estimating the total throughput time of print jobs in the lithography industry?*

In order to address this question printer data needed to be collected first. Approximately two-third was used to calculate three statistical formulas, that each generated a prediction on throughput time for a set of print jobs. In addition, two alternative formulas were included that were frequently brought up by participants. The data that was set aside served for cross-validation of the formulas and for the design of an experimental task which was used to assess participants on their ability to predict throughput time. A comparison of both the answer from the participants against those generated by the formulas, against three criteria, yielded an answer on the research question. To control for an effect of feedback, three different groups of participants took part in this study. That is, *students* who served as a baseline because they are unfamiliar with printing; *operators* who are used to get adequate feedback on the outcome of their predictions and *managers* who served as a control for checking whether frequent and direct feedback is effective. Every participant was assessed on his or her ability to make a prediction for a set of print jobs (i.e. vignette) from which the actual throughput time was known by the researcher.

Disconfirming our most important research question results indicated that operators are outperformed by simple statistical formulas, including a common used heuristic. A possible explanation for this comes from the fact that difficulties were encountered in recruiting operators who had more than 10 years of work experience. Instead, to get the required number of participants involved this criterion was lowered to an average of 10 years. In line with earlier research findings from Ericsson et al. (1993), it is argued that optimal performance could not have been attained because the more experienced (older) operators were underrepresented in our sample. Hence, a more thorough understanding of the data was necessary in order to strengthen this claim.

Moreover, perhaps even more striking with our expectations is that the results showed that managers performed similar to operators. Based on the literature review it was argued that people who get frequent and direct feedback on the outcomes of their predictions should be able to outperform their younger colleagues. If this proposition holds, then it would be particularly interesting to test whether the most experienced operators are able to outperform a statistical formula. Hence, the following sub research question was formulated:

> *[RQ 1.1]Does frequent and direct feedback allow more experienced operators to outperform their younger colleagues, when predicting throughput time for a set of print jobs?*

Confirming this sub research question, results indicated that operators benefit from feedback in a way that more experienced operators outperform their younger colleagues. Moreover, in a similar study that was conducted in the field of purchasing, it was concluded that managers performed worse with growing experience (Snijders et al., 2003). The researchers in this study partly attributed to a lack of direct and frequent feedback. Likewise, as this effect was not found for the group of managers that participated in this study, it seems plausible that these results were caused by a lack of frequent and direct feedback.

In an extra attempt to reject the main research question, a comparison between the predictions from the most experienced operators to those as generated by the different formulas was conducted. Results showed that the accuracy of the predictions improved considerably as a function of work experience, but that even the most experienced group of operators was not able to outperform a four-predictor linear model.

As it turns out that frequent and direct feedback is effective, one last sub question was formulated. In line with earlier research findings from Early et al. (1990), it was argued that individuals benefit from feedback in way that discrepancies between predefined goals and the outcome of one's prediction is used as the basis for self-reflection and consequently for improving performance. If this proposition holds for operators, then it would be particularly interesting to test whether the most self-confident operators are able to outperform a statistical formula. Hence, the following sub research question was formulated:

> *[RQ1.2]Does frequent and direct feedback allow more self-confidence operators to outperform their less self-confident colleagues, when predicting throughput time for a set of print jobs?*

Confirming this sub research question, results indicated that more self-confident operators, indeed, outperform their less experienced colleagues. However, we did not find this result for managers and therefore this finding can also be attributed to a positive effect of frequent and direct feedback.

Moreover, a comparison between the predictions from the most self-confident operators to those as generated by the different formulas was conducted. Although performance slightly improved as a function of an increase in self-confidence, it was concluded that even the most self-confident operators were not able to outperform a four-predictor linear model.

Overall, this leads us to conclude that the results, presented so far, are in favor of the statistical approach. However, before arriving at the main conclusion of this research, it was aimed to rule out one last possibility. That is, given the probabilistic nature the linear models, it was tested how well they perform on specific parts of the data. Hence, it might be that operators outperform the formulas for unusual situations as for example when *[1] print capacity decreases or [2] down time increases*. Note that for the comparison between operators and the different formula, the total group of operators was used.

Confirming this expectation, results indicated that the accuracy of the four-predictor linear model was, indeed, affected by it. Interestingly, the average group operators now reached the performance of the four-predictor linear model but both were outperformed by a commonly used heuristic. In line with earlier research findings from Shanteau (1992), it is concluded that this finding shows that heuristics can be highly accurate under highly specific conditions but also why they are so easily misused.

With this last exception, we conclude that attending operators are outperformed by a four-predictor model when predicting throughput time for a set of print jobs. In line with earlier research findings from Grove and Meehl (1996), the massive opportunity to learn better judgment practices did not result in experienced operators, with more than a decade of work experience, doing nearly as well as a four-predictor linear model. In the next section I elaborate on these findings and compare them to those of related research and try to account for any discrepancies between them.

5.1 Discussion

The general conclusion of this research is that machine printing operators, who are used to get frequent and direct feedback on the outcomes of their predictions, are not able to outperform a four-predictor linear model. In this section, three arguments are provided to consolidate the findings as brought up by this research. Subsequently, two main arguments are provided that indicate serious limitations of this study.

In line with earlier research findings from Grove and Meehl (1996), we have shown that feedback, indeed, improves individuals' performance. However, it only moves younger (less experienced) operators closer to the more experienced ones but does not enable the latter to outperform a statistical formula. A reason why statistical predictions are favored in general is that often a large number of attributes need to be incorporated, and often it is not immediately obvious which attributes have the largest influence and experience is considered to play an important role (Snijders et al., 2003). In our context, however, there were only five attributes that had to be incorporated for a prediction on throughput time. Moreover, operators were (highly) familiar with the task so that they knew what information was relevant and what was not. One argument that remains, therefore, is the role of experience. This is indeed an aspect that was addressed in the analysis but it was not found that more experienced operators outperformed the statistical formulas. Moreover, our results suggest that individuals even cannot outperform statistical models on mundane tasks such as a prediction on throughput time for a set of print jobs. An argument that makes our results even more interesting is that in the lithographic industry, similar conditions arise from day-to-day and therefore offers great advantageous for learning.

A second argument that consolidates this conclusion is that in the majority of experiments in the field of 'clinical versus statistical' prediction, experts often are 'forced' to make their decision in a context that is certainly much more abstract than they are used to (cf. Tazelaar & Snijders, 2004). However, the experimental task was specially designed so that it was realistic and well-known by operators. Given that our conclusion still holds if we only use the data of the operators that were extremely confident in their predictions, we have shown that even for mundane tasks individuals do not outperform a four-predictor linear model.

Lastly, results from our analysis of how participants arrived on their answers provide additional support for the main conclusion. Apparently, students and less experienced professionals (operators and managers) described their method in great detail, whereas the most experienced operators left this question open. This observation suggests that the more experienced operators have difficulties verbalizing what method they are using and that they know more as what they say or express. In line with the research findings from Arts (2007), this condensed communications style is attributed to a higher level of expertise that is often prevalent in more experienced individuals. For example, in a study about the development of expertise in solving mathematical problems, it was found that initially, the participants wrote the entire original formula down, but when expertise level increased, they only noted the application and outcomes of the formula (Sweller, Mawer, & Ward, 1983). Such research illustrates that higher expertise levels use their knowledge in a more tacit or implicit way then novices (De Jong & Ferguson, 1996). Taken together, this implies that the most experienced operators that took part in this study, indeed, possess higher levels of expertise. However, the results of a comparison between the most experienced operators and the different formulas yielded a result that favored the four-predictor linear model.

An argument that restricts the conclusion of this research is that printing machine operators have no real incentive to make an accurate decision. As it turned out, the consequences of an inaccurate prediction in the lithographic industry are not severe. Usually this results in overtime work or, more sporadic, a warning from the higher management. In addition, inaccuracies often are attributed to external sources, such as mechanical breakdowns or a paper type that was not on hand. On the contrary, Wright et al. (1996) have shown that for more responsible professions, individuals are able to outperform a computerized prediction. That is, it was investigated whether attending surgeons could provide more accurate estimates of operating times and compared them against those generated by commercial scheduling software. Interestingly, their results favored the predictions of the surgeons. For an explanation, it was argued by the researchers that inaccurate surgical scheduling influences other departments as for example, recovery room, intensive care, and nursing ward. Moreover, surgeons are kept responsible for delays, especially because procedures are highly standardized and therefore making it unlikely that work is disrupted by external factors. Taken together, this research illustrates that domains where the consequence of inaccurate prediction are large and can be attributed to internal failure are interesting to consider in future research. However, since it was not mentioned what method was underlying the scheduling software we cannot fully strengthen this claim. Nonetheless, it can still be that the scheduling software performed poorly.

A second counterargument that can be made against the results is that operators had to make a prediction on throughput time for a set of print jobs that was actually printed by someone else. Button and Sharrock (2002) found that identifiable individuals differ from others in respect of their printing skills, their initiative, and their reliability. Consequently, a printer will print the same amount in a shorter time, it was argued. This conclusion shows the importance of interpersonal difference when making a prediction on throughput time. In this study, however, the printer data that was used to calculate the formulas and to design the experimental task, originated from a process where four operators were employed. When establishing the linear models, it was already notified that some important predictor was missing. That is, the best linear model only accounted for 80 percent of the variance that was in the data. Given that printing is a mechanical process, it was expected that this value would have been around 90 percent. However, all attributes were included in the linear model, it is argued that the rest of the variance may be explained by interpersonal differences (which was also brought in interviews with operators). However, as the statistical models and the experimental task were established from the printer dataset, both conditions (participants and formulas) were affected by it. In any case, the point remains that, apparently, frequent and direct feedback does not allow operators to outperform a four-predictor linear model.

5.2 Scientific implications

The results as put forward by this research have a number of implications for the research domain that is associated with ‘clinical versus statistical’ prediction. First of all, with this study we have shown that printing machine operators benefit from frequent and direct feedback on the accuracy of their prediction. However, despite this massive opportunity to learn better prediction practices did not result in operators’ doing nearly as well as a four-predictor linear model and replicates a result brought up by an excessive literature review (cf. Grove et al., 2000, p. 20). More in particular, the lithographic industry and the nature of the experimental task would have been classified by several researchers in favor of the decision maker. First, Shanteau (1992) argued that humans show better performance than novices for environments that involve static processes or objects and in which the task people face is highly repetitive. Indeed, we showed that more experienced operators perform better than their younger colleagues do but that they are not able to surpass a four-predictor linear model. Secondly, Snijders et al. (2003) described that statistical models general do well in environments where a large number of attributes need to be incorporated (i.e. typically more than 5) and where it is not immediately obvious how to determine the individual impact of the predictions. Subsequently, it was argued that the tasks participants have to carry out are often too abstract and therefore it becomes impossible to make an optimal prediction(Tazelaar & Snijders, 2004). For this research, an environment was selected that addresses these limitations and designed an experimental task that was easy to administer, understandable and machine printing operators were all familiar with. Nonetheless, we must conclude that simple statistical models beat human expertise in environments that are designated in the literature as predictable and beneficial for human decision makers.

5.3 Practical implications

The results as put forward by this research imply the superiority of a four-predictor linear model instead of relying on operators with more than a decade of experience, for a prediction on throughput time for a set of print jobs. Let us first stress that this is definitely true for managers, whereas this is somewhat different for experienced operators. That is, the results of our analysis showed that managers may not have been able to reflect upon their behavior due to a lack of frequent and direct feedback. Consequently, this restricts them to improve their performance to make accurate predictions. Instead, they tend to make use of external resources (e.g. talking to operators or the manufacture’s specification). Hence, we believe that they are potential candidates for computerized decision-making. Such a tool not only is a credible source, it also enables them to make legitimate decisions based on the actual print capacity instead of the nominal capacity. Though strongly underrepresented in our sample, more experienced operators seem to be quite proficient in this. Nonetheless, we suggest allowing senior operator guidance at the workplace to provide young employees with *feedback* and *reflection* and to use a *statistical model* to make a prediction on *throughput time*.

More general, operators might perceive the implementation of such a tool as a threat to their autonomy rather than as an enrichment to their work (Hackman et al., 1975). To cope successfully with such an intervention in the workplace, it is important to have a notion of how such a tool affects the psychological states of workers. That is, to increase the motivation and satisfaction of workers, a worker must believe that he personally is accountable for the outcome of his or her efforts, also referred to as experienced responsibility or autonomy (Hackman et al., 1975). Due to the ever-increasing computerization of nowadays industrialization, this imperative is easily violated. Our method accounts for this by giving the operator freedom, independence, and discretion in scheduling work. That is, we see our method being implemented as a tool that generates a target time that is realistic and process specific. By doing so, we introduce a more direct form of outcome feedback. In addition, various studies have shown that goal-setting enhances performance through increasing effort and persistence, directing attention, and improving strategy formulation (cf. Early et al., 1990). Motivationally this helps for instilling a sense of competence, accomplishment and control in workers, it is argued. Accordingly, Hackman et al. (1975) stated that job-provided feedback is usually more immediate and discrete than supervisor-supplied feedback, which contributes to the operator' perception of control. In this way, he frequently gets feedback on his performance and does not depend on messages or subtle cues from the manager. We also see the opportunity to *'praise'* operators with a positive feedback message when a period of error-free performance has been sustained and by doing so increase its persuasive potential to support printing machine operators to work more efficiently (Fogg, 2003).

Note that we do not claim that multiple linear regression analysis is the best method to predict throughput time for a set print job. There likely are methods that are more reliable and accurate or more straightforward to use. However, based on findings of this research we hope to put forward two things. First, we conclude that work activities are inherently *'noisy'* and therefore there is absolutely no need for a method that is extremely accurate or provides a *'fixed'* working schedule. Typically, high priority jobs arriving unexpectedly or jobs require reprinting because of a mistake. Our method deals with this by showing the impact of unexpected work on the overall planning while the operator remains responsible to cope with this extra workload. In addition, a manager can anticipate upcoming delays and decide if overtime working is needed or that a client needs to get informed. Secondly, we have shown that a four-predictor linear model predicts throughput time for a set of print jobs better than a print machine operator with a decade of work experience. In addition, implementation of this method is relatively straightforward since most of the necessary functionality is already there. Using relatively little data it has been shown that an accurate formula can be established that performs better than an operator with more than 10 years of work experience. However, as long as this is not reality, we recommend listening to experienced operators for a prediction on throughput time.

5.4 Limitations

As with any study, this research was subjected to limitations. Let us first emphasize that this research was conducted as field experiment and therefore it was more difficult to control for extraneous variables. We discuss the ones that directly relate to the design of the experimental study. First, at larger companies operators worked in shifts which made it impossible to test all operators on the same day. As a result, we cannot rule out the possibility that the purpose of the study became a topic of conversation in the intervening days and that operators were influenced by one and another. Secondly, from a financial point of view some companies printed on A3 paper only. Although this slows down the process substantially, it might be that the operators were influenced by this. However, we cannot rule out this possibility since the experimental task was completed anonymously. Lastly, we address three possible threats to the validity of our results that we hope to address in a future experiment.

- Printing machine operators with more than 10 years of work experience were underrepresented in our sample. As performance increases as function of work experience, we believe that this is a serious threat the results obtained. Hence, we recommend only to included older (senior) operators in future studies.
- Ericsson (2004) state that more years of experience do not automatically guarantee that one reaches optimal performance, but rather the quality of the experience is important. Hence, we observed that operators controlled several printers (from different types and manufactures) simultaneously. This suggests that operators did not benefit as much from the feedback as when they were in a one-to-one relationship with a printer. This is something to account for in a future experiment.
- The large variability in print processes made it difficult to define a task that was generalizable. As a result, it might be that operators' predictions were based different process characteristic (e.g. print capacity) and consequently scored worse.

5.5 Future directions

With this research we have shown that there is much to improve and that calculations based on knowledge from personal experiences, a limited number of attributes and for well-known tasks are outperformed by relatively simple formulas. Note that the focus in this study was on printing only and that planning the whole process (i.e. from job intake to shipping) remains to be a challenging endeavor. For example, Button & Sharrock (2002) describe the three most common contingency attributes that are considered to be important for planning work. That are:

- The *customer* who has the reputation to change their mind at the last minute, deliver materials late or that their jobs contain unexpectedly demanding characteristics that generate additional work from a print point of view (customers often are therefore reputed as being ‘bad’ or ‘good’).
- The *organizational processes* which depends upon on a number of departments or ‘steps’ (i.e. job intake, preparation, printing, finishing and shipping) that have been especially designed to ensure that the print process is divided into different stages (e.g. pre-press, press, post-press) and workload is easier to oversee.
- The *shop floor* deals with the production of an order and the fine differences between operator and printers.

Taken together, in a future experiment we hope to address whether a single computerized decision-making tool is able to account for all the above stated aspects. Interestingly, a method proved already to be successful in more complex environments, involves combining the prediction of both the expert and the model to get a ‘model-plus-expert’ prediction (cf. Snijders et al., 2003). The general idea why this works is in line with the results we have found so far, that is, experts generally deviate from linear predictions in a strong manner. In this way, the model takes care of an overall linear fit that is reasonable, whereas the expert’s intuition could capture the deviations from the linear model.

References

Button, G., & Sharrock, W. (1997). *The production of order and the order of production: possibilities for distributed organizations, work and technology in the print industry*. Paper presented at the Proceedings of the fifth conference on European Conference on Computer-Supported Cooperative Work.

Button, G., & Sharrock, W. (2002). Operating the production calculus: ordering a production system in the print industry. *British Journal of Sociology, 53*(2), 275-289.

De Jong, T., & Ferguson, M. G. M. (1996). Types and qualities of knowledge. *Educational Psychologist, 31*(2), 105-113.

Dexter, F., Traub, D. R., & Qian, D. (1998). Comparison of statistical methods to predict the time to complete a series of surgical cases. *Journal of Clinical Monitoring and Computing, 15*, 45-51.

Early, P. C., Northcraft, G. B., Lee, C., & Lituchy, T. R. (1990). Impact of process and outcome feedback on the relation of goal setting to task performance. *Academy of Management Journal, 33*(1), 87-105.

Ericsson, A. K., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363-406.

Ericsson, A. K., & Lehmann, A. C. (1996). Expert and exceptional performance evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47*(1), 273-305.

Field, A. (2005). *Discovering Statistics using SPSS*. London: SAGE Publications Ltd.

Fogg, B. J. (2003). *Persuasive Technology*. San Francisco: Morgan Kaufmann Publishers.

Galtung, J. (2003). What did the expert predict? *Futures, 35*, 123-145.

Graham, J. R. (1971). Feedback and accuracy of clinical judgments from the MMPI. *Journal of Consulting and Clinical Psychology, 36*(2), 286-291.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical-ctatistical controversy *Psychology, Public and Law, 2*(293-323).

Grove, W. M., Meehl, P. E., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological Assessment, 12*(1), 19-30.

Hackman, R. J., Oldham, G., Janson, R., & Purdy, K. (1975). A new strategy for job enrichment. *California Management Review, 17*(4), 57-71.

Mandela, N. (1995). *Long walk to freedom: The Autobiography of Nelson Mandela*. Great Britain: Abacus.

Rai, S., Duke, C. B., Lowe, V., Quan-Trotter, C., & Scheermesser, T. (2009). LDP lean document production-O.R. enhanced productivity improvements for the printing industry. *Interfaces, 39*(1), 69-90.

Richins, S., & Holmes, M. (1998). Waiting for satisfaction. *Journal of Healthcare Management, 43*(3), 281-286.

Schroeder, H. E. (1972). Use of feedback in clinical prediction. *Journal of Consulting and Clinical Psychology, 38*(2), 265-269.

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes, 53*(2), 252-266.

Snijders, C., Tazelaar, F., & Batenburg, R. (2003). Electronic decision support for procurement management: evidence on whether computers can make better procurement decisions. *Journal of Purchasing & Supply Management, 9*, 191-198.

Sweller, J., Mawer, R., & Ward, M. (1983). Development of expertise in mathematical problem solving. *Journal of Experimental Psychology, 112*(4), 639-661.

Tazelaar, F., & Snijders, C. (2004). The myth of purchasing professionals' expertise. More evidence on whether computers can make better procurement decisions. *Journal of Purchasing & Supply Management, 10*, 211-222.

Wright, I. H., Kooperberg, C., Bonar, B., & Bashein, G. (1996). Statistical modeling to predict elective surgery time: comparison with a computer scheduling system and surgeon-provided estimates. *Anesthesiology, 85*(6), 1235-1245.

Appendices

A - Properties of the statistical models that were established

Four-predictor linear model with total number of jobs, papertraus and one-sided and two-sided sheets as a predictor

Total throughput time =8.256 + (0.497*jobs) + (3.44*total_papertrays) + (0.006*one-sidedsheets) + (0.004*(2*-two-sidedsheets))

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,890[a]	,792	,767	61,41191	,792	31,443	4	33	,000	2,144

a. Predictors: (Constant), total number of jobs, 2 sided prints, total number of papertrays, 1 sided prints

b. Dependent Variable: Printtime_minutes

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	474346,299	4	118586,575	31,443	,000[a]
	Residual	124456,964	33	3771,423		
	Total	598803,263	37			

a. Predictors: (Constant), total number of jobs, 2 sided prints, total number of papertrays, 1 sided prints

b. Dependent Variable: Printtime_minutes

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	8,256	36,477		,226	,822					
	1 sided prints	,006	,002	,282	3,073	,004	,314	,472	,244	,749	1,336
	2 sided prints	,008	,001	,745	9,159	,000	,710	,847	,727	,952	1,051
	total number of papertrays	3,440	1,250	,233	2,752	,010	,345	,432	,218	,876	1,141
	total number of jobs	,497	,231	,204	2,155	,039	,464	,351	,171	,704	1,420

a. Dependent Variable: Printtime_minutes

One-predictor linear model with the total number of two-sided sheets as a predictor

Total throughput time =180.194 + (0.008*two-sidedsheets)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,710[a]	,504	,490	90,86739	,504	36,522	1	36	,000	1,970

a. Predictors: (Constant), 2 sided prints

b. Dependent Variable: Printtime_minutes

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	301555,512	1	301555,512	36,522	,000[a]
	Residual	297247,751	36	8256,882		
	Total	598803,263	37			

a. Predictors: (Constant), 2 sided prints

b. Dependent Variable: Printtime_minutes

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	180,184	37,198		4,844	,000					
	2 sided prints	,008	,001	,710	6,043	,000	,710	,710	,710	1,000	1,000

a. Dependent Variable: Printtime_minutes

One-predictor linear model with the total number of impressions as a predictor

Total throughput time = 120.439 + (0.004*Impressions).

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,796[a]	,633	,623	78,11415	,633	62,135	1	36	,000	1,925

a. Predictors: (Constant), total number of prints

b. Dependent Variable: Printtime_minutes

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	379137,709	1	379137,709	62,135	,000[a]
	Residual	219665,554	36	6101,821		
	Total	598803,263	37			

a. Predictors: (Constant), total number of prints

b. Dependent Variable: Printtime_minutes

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	120,439	36,063		3,340	,002					
	total number of prints	,004	,001	,796	7,883	,000	,796	,796	,796	1,000	1,000

a. Dependent Variable: Printtime_minutes

B: Vignette Experiment / demographics

Bent u man of vrouw?   M   -   V

Wat is uw beroep?

Wat is uw leeftijd?

1. Wat is u functie?

Student	Voorbereider	Print operator	Finisher	Allround operator	Anders nl:

2. Hoeveel jaar werkervaring hebt u in de print industrie?

Minder dan 1 jaar	1 tot 5 jaar	5 tot 10 jaar	10 tot 15 jaar	15 tot 20 jaar	Meer, nl:

3. Hoeveel uur per week werkt u met de Xerox printer, type Nuvera 288?

Minder dan 10 uur	10 tot 20uur	20 tot 30uur	30 tot 40uur	Meer, nl:

4. Hoeveel jaar / maanden werkt u al met de Xerox printer, type Nuvera 288?

Een paar maanden	Een half jaar	1 jaar	1,5 jaar	2 jaar	Meer dan 2 jaar

5. Hoeveel afdrukken worden er op de printer waar u mee werkt gemiddeld per dag gemaakt?

Minder dan 50.000	50.000 - 100.000	100.000 -150.000	150.000 -200.000	200.000 -250.0000	Meer, nl:

6. Wat is de veRhouding van enkelzijdige en dubbelzijdige afdrukken die gemaakt worden op de Xerox printer, type Nuvera 288?

Enkelzijdig gemiddeld per dag: …………	Dubbelzijdig gemiddeld per dag: ……….

7. Bent u verantwoordelijk voor het registreren van tellerstanden?

Nee			
Ja, ik doe dit:	Dagelijkse	Weekelijks	Maandelijks

8. Maakt u veel gebruik van de online finishing mogelijkheid die de printer biedt?

Nee				
Ja, vooral:	Nieten	Boren	Vouwen	Anders, nl:

C: Background information & Vignette experiment

Stelt u zich nu voor dat u een printshop met 1 printer beheert (zie Figuur 11). Iedere ochtend staat voor u een overzicht klaar hoeveel en wat voor soort print opdrachten die dag dienen te worden verwerkt. Straks is het uw taak om op basis van de aangegeven set van print opdrachten een inschatting te geven of te berekenen hoe lang het duurt voordat het totaal aan opdrachten is afgedrukt.

U krijgt nu eerst informatie van de printer gepresenteerd.

- **288** dubbelzijdige of **144** enkelzijdige **afdrukken per minuut**
- Het laadvolume van de papierladen bedraagt **2x 5800** vel en **4x 3200** vel A4.
- Zaken waar de bediende geen controle over heeft, zoals **storingen**, **vastlopend papier** e.d. kunnen de tijdsduur negatief beïnvloeden.
- De bediening geschiedt zo **optimaal** mogelijk zodat soortgelijke opdrachten veelal continue kunnen worden geprint.

Figure 11 Xerox Nuvera 288

U krijgt nu eerst informatie van de opdrachten gepresenteerd.

- Het merendeel van de opdrachten wordt afgedrukt op **75 grams** papier voor deze opdrachten zijn **maximaal 3 papierladen** benodigd.
- Een klein aantal opdrachten (polissen) bestaan meestal uit een aantal vellen **75 grams** en **90 grams** papier. Voor deze opdrachten zijn minstens 4 papierladen benodigd.

<u>LET OP:</u> er wordt niet gecorrigeerd voor de verschillen in papiergewichten!

Op een dag kunnen opdrachten voor **dezelfde** klant of van **verschillende** klanten worden. Opdrachten **binnen een set** worden dezelfde papiersoort vaak **aangevuld** met een ander type. Voor **elke nieuwe set** dienen alle papierbakken te worden **geleegd** en te worden gevuld met het papier van een andere klant.

Uw taak:

We vragen u om zo meteen om voor een 8-tal opdrachten de totale printtijd voor de betreffende dag in te schatten. Van iedere opdracht krijgt u steeds:

- het aantal sets dat op een dag verwerkt dient te worden.
- het **benodigde papier** en een overzicht van de opdrachten die op dit papier dienen te worden geprint.
- het totaal aantal afdrukken
- totaal aantal **enkelzijdige** en **dubbelzijdige** afdrukken
- totaal benodigde aantal **papierladen**.

Er volgt nu een voorbeeld van een vignet waarvoor print dient te worden geschat (zie Figuur 12). Ieder vignet bevat de volgende informatie:

- Aantal **sets (**soorten verschillende klanten)
- Afdrukken
- Aantal **enkelzijdig** bedrukte **vellen**
- Aantal dubbelzijdig bedrukt vellen
- Maximaal aantal te gebruiken **papierladen**

Let op: de totalen zijn vetgedrukt weergegeven onder de betreffende kolommen.

49

Set 1	Totalen Opdrachten	Afdrukken	Enkelzijdige bedrukte vellen	Dubbelzijdige bedrukte vellen	Te gebruiken papierladen
polispapier	**73**	**72994**	**1650**	**35672**	**4**
430X102	28	7185	459	3363	4
430X97D	15	22924	202	11361	3
430X98D	30	42885	989	20948	3
briefpapier	**2**	**98**	**34**	**32**	**2**
430X085	2	98	34	32	2
Totalen	**75**	**73092**	**1684**	**35704**	**4**

Geschatte benodigde tijdsduur voor alle opdrachten samen (omsteltijd + machinetijd):
(geef een zo precies mogelijke schatting)

.... uur en .... minuten

Hoe goed verwacht u dat uw schatting is (omcirkelen wat van toepassing is):

o Hooguit 10 minuten ernaast
o Hooguit 30 minuten ernaast
o Hooguit 45 minuten ernaast
o Hooguit 1 uur ernaast
o Hooguit 1.5 uur ernaast
o Mogelijk meer dan 1.5 uur ernaast

Figure 12 Overizicht van een set aan print opdrachten

Op basis van uw ervaring dient u een inschatting te geven voor de totaal benodigde tijdsduur voor het printen van een dergelijke opdracht (printtijd en omsteltijd) en aan te geven hoe ver uw schatting afwijkt van de werkelijke tijd.

Denkt u dat u gegeven uw ervaring redelijke inschattingen kunt maken?

- zeker wel (1)
- waarschijnlijk wel (2)
- het kan goed of slecht gaan (3)
- waarschijnlijk niet (4)
- zeker niet (5)

Gegeven uw ervaring, wat is volgens u een acceptabele afwijking van uw schatting ten opzichte van de werkelijke tijdsduur? (omcirkelen wat van toepassing is)

0-10%	10-20%	20-30%
30-40%	40-50%	50-60%
60-70%	meer dan 70%	

Er volgen nu een 8-tal opdrachten waarvan u wordt gevraagd een inschatting voor de totale tijd dient te geven.

D: Section to write down what method participants were using

Hartelijk dank voor uw medewerking.

Kunt u hier nu (duidelijk) vermelden welke methode / informatie u hebt gebruikt voor het maken van uw inschatting?

Denkt u dat u gegeven uw ervaring redelijke inschattingen kunt maken?

- zeker wel (1)
- waarschijnlijk wel (2)
- het kan goed of slecht gaan (3)
- waarschijnlijk niet (4)
- zeker niet (5)

Hoe groot schat u uw eigen afwijking tov de werkelijke tijden in? (omcirkelen wat van toepassing is)

0-10%	10-20%	20-30%
30-40%	40-50%	50-60%
60-70%	meer dan 70%	

E: Comparison between groups and models

The tables are designed so that the participant's score (i.e. median) on a criterion is directly viewed in contrast with the model's score (paired). The test statistics are placed right underneath it so that it can easily be observed if the difference is significant or not.

Rank correlation	P	Heuristic	P	LM impressions	P	LM two-sided	P	LM 4-predictors	P	Nominal
Score	,71	,88	,71	,91	,71	,69	,71	,88	,71	,83
Wilcoxon Signed Ranks test	*Z = -4,968 p =,000*		*Z = -6,075 p =,000*		*Z = -,996 p =,319*		*Z = -4,968 p =,000*		*Z = -3,541 p = ,000*	

Table 6 Paired comparison between the group and the different models

Relative deviation	P	Heuristic	P	LM impressions	P	LM two-sided	P	LM 4-predictors	P	Nominal
Students	,33	,14	,33	,13	,33	,20	,33	,13	,33	,33
Wilcoxon Signed Ranks test	*Z = -4,349 p =,000*		*Z = -4,554 p =,000*		*Z = -3,939 p =,000*		*Z = -4,577 p =,000*		*Z = -,751 p = ,452*	
Professionals	,26	,14	,26	,13	,26	,20	,26	,13	,26	,33
Wilcoxon Signed Ranks test	*Z = -4,104 p = ,000*		*Z = -5,098 p = ,000*		*Z = -,1,682 p = ,092*		*Z = - 4,677 p = ,000*		*Z = -3,720 p = ,006*	

Table 7 Paired comparison between groups and models on the score for the relative deviation

Vignettes correct	P	Heuristic	P	LM impressions	P	LM two-sided	P	LM 4-predictors	P	Nominal
Students	1	3	1	5	1	3	1	5	1	0
Wilcoxon Signed Ranks test	*Z = -3,59 p =,000*		*Z = -4,534 p=,000*		*Z = -2,823 p =,005*		*Z = -4,477 p =,000*		*Z = -3,97 p = ,000*	
Professionals	3	3	3	5	3	3	3	5	3	0
Wilcoxon Signed Ranks test	*Z = -,167 p = ,867*		*Z = -4,748 p = ,000*		*Z = -1,599 p = ,110*		*Z = - 3,788 p = ,000*		*Z = -5,184 p = ,000*	

Table 8 Paired comparison between groups and models on the score for the number of vignettes predicted correctly