Measuring consumer emotions in failure situations

Bergmans, R.F.M.

Award date:
2007

Eindhoven, August 2007

Measuring consumer emotions in failure situations

by
Rick F. M. Bergmans

B.Sc Industrial Engineering and Management Science – TU/e (2005)
Student identity number 0536599

In partial fulfillment of the requirements for the degree of

Master of Science
In Operations Management and Logistics

Supervisors:
Prof. Dr. Ir. A.C. Brombacher, TU/e, Quality and Reliability Engineering
Dr. Ir. P.J.M. Sonnemans, TU/e, Quality and Reliability Engineering

Advisory member
Ir. I.M. de Visser PhD-candidate, TU/e, Department of Industrial Design

TUE. Department Technology Management.
Series Master Theses Operations Management and Logistics, nr.6

ARW 2006 OML / 6

Subject headings: consumer research, usability engineering, irritation, emotion

Abstract

This thesis explores the perceived irritation (User Perceived Failure Severity) caused by a product failure in a high end consumer electronic product. Several attributes of a failure that influence UPFS are examined. Problems in the field of reliability engineering are examined by a user centered approach. Therefore emotions research is used to deal with reliability problems. By means of a critical theoretical investigation instruments are selected (and rejected) for measuring UPFS. A model that determines the failure severity from a user perspective and the measurement instruments for UPFS are tested by a quantitative approach, namely two experiments (N=25 and N=149). A validated UPFS model is presented.

Keywords: consumer research, usability engineering, irritation, emotion.

Acknowledgements

This thesis is the end product of my graduation project of the master Operations Management and Logistics. Actually, it can be seen as the final product of the study Industrial Engineering and Management Science at the Eindhoven University of Technology. This project has been carried out at the sub department Quality and Reliability Engineering.

First, I want to thank my supervisors for their input and feedback. Unfortunately Lu Yuan is not on the cover, she is on a pregnancy leave. Nevertheless I would like to thank her for coaching me during my master thesis preparation and helping me with setting the outlines for my thesis project. I would like to thank my first supervisor Aarnout Brombacher for his enthusiasm during this project and for practical issues like building the "signal-disturbance device". Also thanks to Peter Sonnemans for his useful comments during the midterm presentation and in the end phase of this master project.

Special thanks to Ilse de Visser for her commitment with my master thesis project. Her feedback was really valuable. Her critical view on structure and content helped me to improve the quality of this thesis. She helped me with the scientific depth of my thesis and to verbalize unambiguously. Also I would thank her for helping me with the execution of the experiments and for arranging nice presents such as music boxes and a television for the participants.

Also thanks to Bas Gielen whose research with regard to "function importance" was indispensable for my thesis. Together we developed and conducted the first experiment. We could easily converse about difficult content related things, since his research project is closely linked to my research project. This was very helpful.

As fourth I would thank my family and friends for supporting me during these months of graduation and moreover the whole five years of my university life. Last but not least I would thank Rob and Armand who are my flat-mates. Without them these years in Eindhoven would not be as fun and interesting as they have been.

Summary

Manufactures of consumer electronic products are in a continuous fight for performance and technical innovation on the one hand and on the other hand product quality and reliability (Brombacher, Sander, Sonnemans and Rouvroye, 2005). There are four business trends within the consumer electronic industry: 1) Globalization and outsourcing of the business processes, 2) Increasing time-to-market pressure, 3) Increasing product complexity and 4) Increasing customer requirements (Brombacher and De Graef, 2001).

The current business trends cause an information gap between the developers and users of consumer electronic products (Brombacher et al., 2005). There is increased attention for consumer satisfaction combined with a stronger time-to-market pressure. The quality of the feedback from the field does not have the quality for finding root causes of failures and non-technical problems. Due to the lack of this information developers can only use their own technical expertise in failure prioritization. This technical view differs from the user's view. The user perspective is lacking in failure prioritization (De Visser, 2006). Therefore a model should be developed that puts forward the user perspective in the failure prioritization process (Den Ouden, 2006). This is translated into a PhD research with the following research question:

How to develop a validated classification model for the determination of the failure severity from a user perspective?

The accompanying research model is given below:

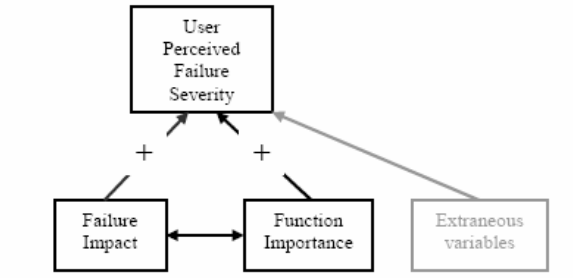


Figure 1: UPFS research model

The User Perceived Failure Severity (UPFS) is the level of user-irritation caused by a product failure. Failure Impact is the percentage loss of functionality as a result of the failure. Function Importance is the relative importance of the function affected by the failure. One important extraneous variable is irritableness. Irritableness can be seen as how quickly people get irritated. Some people get irritated by minor product failures other people get irritated by major product failures.

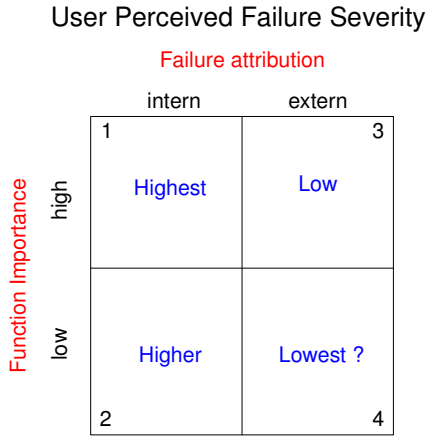
Two experiments were conducted for measuring User Perceived Failure Severity (UPFS) and testing the UPFS research model. UPFS is the level of user-irritation caused by a product failure. Irritation is an emotion (reaction). Emotion reactions consist of stable traits and fluctuating affective transitory states. Stable traits describe how people usually or typically are. Affective states describe the affective condition of people during a specific moment in time. Affect refers to the feeling side of consciousness, like pleasure and displeasure, happiness and sadness, liking and disliking. Irritation is also an emotion.

There needs to be controlled for "irritableness" in consumer research that measure irritation caused by a product failure. Irritableness can be seen as how quickly people get irritated. This level of irritableness

determines the irritation reaction of users confronted with a failure in a consumer electronic product. Users may overreact or "underreact". Therefore users need to be selected with an average normal level of irritableness. A scale that could measure irritableness has to fulfill certain criteria. These criteria are that the scale needs to measure the personal trait: irritableness. The scale has to be valid and reliable. The scale should be easy enough to use for the participant and the researcher. Based on these criteria, PANAS(-X) is suitable for measuring irritableness. PANAS-X is the expanded form of PANAS. PANAS is the Positive and Negative Affect Schedule, which can measure emotional experience. For this application PANAS-(X) was translated into Dutch. Based on an analysis of 200 Dutch PANAS-X questionnaires can be concluded that the use of the original PANAS can control for irritableness. The scale Negative Affect of PANAS is a measurement-instrument for irritableness. Based on the results of the two experiments (N=348), Dutch students with a Negative Affect level between 13.43 and 23.95 have a normal irritableness level. Results confirm the validity and reliability of the Dutch version of PANAS.

Also some conclusion can be drawn about the Dutch version of the eXpanded form of PANAS. The reliability of the scales of the Dutch version of PANAS-X are reasonably well. Especially the scales related to Negative and Positive Affect. Two scales (Shyness and Serenity) measuring "other affective" states don't posses the minimal acceptable reliability level. The reliability of the scale Shyness is low and the reliability of Serenity is unacceptable low. With regard to validity can be said that the basic positive emotions scale (Joviality, Self-Assurance and Attentiveness) have sufficient convergent and discriminant validity. The basic negative emotion scales (Fear, Hostility, Guilt and Sadness) have sufficient convergent validity but there is a lack of discriminant validity. This means that it is not possible to specify the Negative Affect Emotions into Fear, Hostility, Guilt, or Sadness.

For measuring UPFS a scale is needed that measures irritation. For measuring UPFS not the stable traits (as were the case for irritableness) are at interest but the affective transitory states. Besides this criterion the measurement instrument needs to be valid and reliable and also easy to use for participants and researchers. Based on the results of the two experiments, the irritation caused by a failure in a consumer electronic product (UPFS) can be measured with the Emotional Response Scale. This scale is reliable and valid and easy to use. PANAS-X can not measure UPFS. The reason is that PANAS-X can not detect differences between a "normal situation" and a failure situation well. However the Emotional Response Scale can measure UPFS, is valid and reliable and easy enough to use for participants as well as for the researchers.



The result of the first experiment (N=25) is that besides failure importance: the relative importance of the function affected by the failure, also failure attribution plays an important role in UPFS. Failure attribution means whether the cause is internally attributed or externally attributed. Internal attribution is that the cause of a failure is attributed to the participant or the television itself. External are all things beyond internal, like blaming the weather, the cable company or the power company for a failure. The level of UPS for these different scenarios is visualized in figure 2. This hypothetical UPFS-matrix is based on quantitative and qualitative results of the first experiment.

Figure 2: UPFS-matrix

In the second experiment the UPFS matrix was tested (N=149). Four different scenarios were tested and the UPFS-levels of these four scenarios were compared. The numbers in the UPFS-matrix represent the

different scenarios. The highest UPFS occurs in a situation where a high importance function internally attributed fails (scenario 1). This UPFS is significant different from the situation where a low importance function externally attributed fails (scenario 4), although not significant different from the other scenarios (scenario 2 and 3). The UPFS of a low importance function internally attributed failure (scenario 2) is higher than the UPFS of a high importance function externally attributed failure (scenario 3). However the difference is not significant.

The UPFS-level is influenced significantly by Failure Attribution (FA) and perceived Failure Impact (FIQ). Failure Impact is the loss of functionality as a result of the failure. When the Failure Attribution moves from extern to intern UPFS increases. When the Failure Impact increases the UPFS also increases. Function Importance (FI) is not significant in the second experiment but is significant in the first experiment. An explanation could be that in the second experiment participants are asked to evaluate the failures. In the first experiment people are confronted with a failure while doing a task. The emotional reaction in the second experiment is less than in the first experiment. Therefore the differences found in function importance in the second experiment are not significant but are nevertheless in the proposed direction. Based on the results of both experiment the UPFS end-model can be presented, figure 3. Thicker lines stand for larger effects on UPFS.

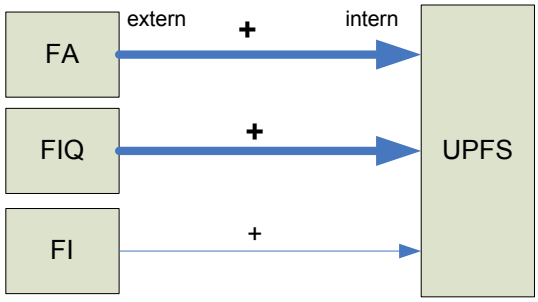


Figure 3: UPFS end-model

The obvious results with regard to failure impact imply that a larger failure impact will lead to a higher UPFS-level. More interesting is that despite the fact that all participants were exposed to the same level of failure impact, they perceived this failure impact rather differently and consequently this resulted in a different UPFS-level. It is confirmed that failure attribution influences UPFS (second experiment). It has to be noted that an externally attributed failure which actually is an internal failure will not be perceived externally enduringly. For example, in the situation where there is a failure in the picture you could attribute this failure externally to the cable company. However when this failure occurs several times it is likely that due to social contact (with the neighbors) it becomes clear that your television is the only television which has a problem with the picture. In this situation it is likely that the users will (re)attribute the picture failure internally. From the first experiment it is known that whether the failures are attributed internally or externally is person dependent. Also just as failure impact the results with failure attribution are person specific. Therefore UPFS is largely person dependent. How important a function is also depends on the user. Gielen 2007 concludes that there is moderate agreement between participants on the importance of functions. This is confirmed in this thesis project. For high importance functions there is high agreement on the importance of the function. For low importance functions there is less agreement on the importance of the functions. This can be explained by the fact that less important function are relatively new functions like Ambilight and the motorized swivel. Users need to explore these functions to capture the real added value of these functions and to evaluate these functions whether they are important to them or not.

Recommendations

High-end consumer electronic manufacturers should take into account that UPFS is heavily user dependent. For reliability optimization, users with the same characteristics with regard to failure attribution, failure impact and function importance should be categorized. This means that it is not enough to classify users only based on the Rogers-curve (1983) like early adaptors, or laggards. An extension should be made with the factors listed above. Based on this segmentation of users, different market strategies can be developed. This segmentation means that some functions which are not completely developed and tested, but are from marketing viewpoint very attracting to add, can be included in one segment and excluded in the other segment. Users with a low perceived failure impact level who tend to attribute failures mostly external will perceive a failure less severe than a user with a high perceived failure impact level mostly attributing failures internal. So, for the first group it is attractive from marketing and reliability viewpoint to add this function in the product for this group, since the UPFS will be low when a not completely developed and tested function fails. Based on function importance a prioritization of failures can be made. However, also the importances of new functions are partly user specific. So, resolving a failure will add more value to one segment than to the other segment. Therefore the failure of a function that is seen by the most profitable user group as most important should be resolved first. In conclusion this can be summarized in the figure below.

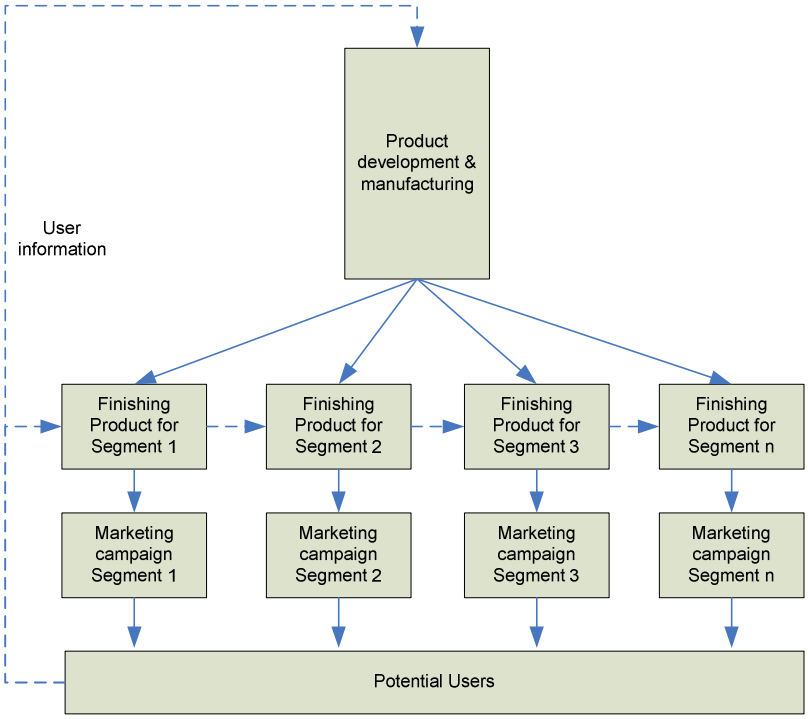


Figure 4: Segmentation by use of the UPFS model

Based on (marketing) research (potential) users are categorized in different segment based on Failure Impact, Failure Attribution and Function Importance. The “base-product” will be adjusted to the specific users segments in order to minimize reliability problems. This means that certain functions are included for some segments but are excluded for other segments. Specific marketing campaigns should take care for the distribution of the specific segmented products to the accompanying (segmented) users.

Abbreviations

ANOVA:	Analysis of variance
AR:	Action Response
ER:	Emotional Response
ERS:	Emotional Response Scale
FA:	Failure Attribution
FAQ	Failure Attribution Questionnaire
FI:	Function Importance
FIQ :	Failure Impact Questionnaire
MSA:	Measure of Sampling Adequacy
NA:	Negative Affect
NFF:	No Failure Found
PA:	Positive Affect
PANAS:	Positive and Negative Affect Schedule
PANAS-X:	Positive and Negative Affect Schedule – eXpanded Form
PhD:	Doctor of Philosophy
SD:	Standard Deviation
UPFS:	User Perceived Failure Severity
VIF:	Variance Inflation Factor

Table of contents

List of Figures

List of Tables

1. Introduction

Manufactures of consumer electronic products are in a continuous fight for performance and technical innovation on the one hand and on the other hand product quality and reliability (Brombacher, Sander, Sonnemans and Rouvroye, 2005). These innovative products with new technologies in combination with customers that are unfamiliar with these technologies, lead to (unpredicted) problems which are hard to manage (Brombacher et al., 2005). There are four business trends identified within the field of high volume consumer electronic products. They are: globalization and outsourcing of the business processes, increasing time to market, increasing product complexity and increasing customer requirements (Brombacher and De Graef, 2001).

Globalization and outsourcing of the business processes
Producers of consumer electronic products opened factories in "low wage" countries, due to increased international competition (Murthy, Kadur and Nagaraju, 1994), (Classen and Lopez, 1998), (Brombacher and De Graef, 2001). Some producers also outsourced several activities, so they can focus entirely on their core business. This outsourcing is going quite far, parts of the development process are outsourced (Dijkstra, Dirne Govers and Sander, 1997), (Petkova, 2003). Also service and repair activities are outsourced for efficiency. When a customer returns his product for repair the customer will get his product back as soon as possible. This results in a situation where there is less focus on gathering root cause data for product improvement (Den Ouden 2006). These outsourcing activities bring some risks with them. Due to globalization, the value chain becomes disintegrated. Information needs to be transferred not only to different disciplines but also via different countries and companies. This may reduce the reliability of the information, loss of data integrity, delay in information flows and loss of information in the chains (Den Ouden, 2006).

Increasing time-to-market pressure
The speed to bring new technology to the market has increased considerately, due to the first-mover advantage (Wheelwright and Clark, 1992), (Foo, Lien, Xie and Van Geest, 1995), (Brooks and Schofield, 1995). The company who enters first the market (first mover) will have a competitive advantage until the second company enters the market. A first mover can set premium prices for his product because there is no competition. High revenues are expected for first-movers. Another issue companies are faced with, when not-being a first mover, is market-saturation. At the time other companies (no first-movers) enter the market some potential customers already bought the product from the competitor. The number of potential customers will decrease and subsequently the expected revenues (market saturation). Therefore, it is important to bring the technology to the market as quick as possible. This results in a time-pressure in the product creation process. Products not being in time on the market, can result in two unfavorable situations: the product has to compete with cheaper products or with products with more functionality (Den Ouden, 2006).

In some cases, due to time pressure some activities are not properly done, or even cancelled. These activities are mostly related to the testing of the product. Inevitable this may lead to uncertain reliability levels (Minderhoud, 1999). This will result in the user sending back the product for repair, or even worse, a re-call of all products of a certain type. A re-call of all products is a very expensive operation. All (re-called) products have to be repaired or substituted. Also, the image of the company who is re-calling the products will be damaged. This could go so far that complete brands and even companies are destroyed (Smith, Thomas and Quelch, 1996).

The development time has been reduced, but the time to get feedback from the field remains almost the same, see figure 1.1(Brombacher et al., 2005). The development time of consumer electronic products is

within a range of 6 to 9 months, while the feedback time is larger than 1 year (figure 1.1). This will lead to situations where information about the product reliability of the first generation is available when the third generation is being developed (Den Ouden, 2006).

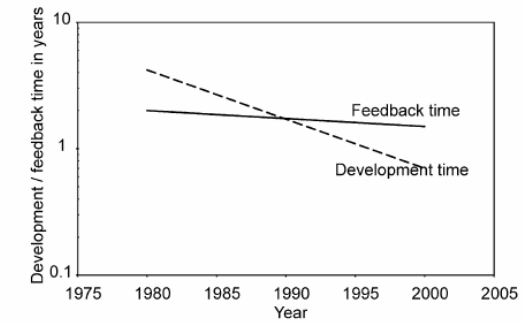


Figure 1.1: Development time versus feedback time for high volume-consumer electronics (Brombacher, 2002)

Increasing product complexity

Increasing product complexity is caused by adding more functionalities to a product. Technologies and functionalities become available at lower prices. These functionalities interact with each other, and if more functionalities are added to products, prediction of product interactions becomes more difficult (Petkova, 2003). Also another point is that more products are involved in networked environments. Products have to communicate with each other (for example a PC with a mobile phone). Therefore Den Ouden also stated that analyzing quality problems becomes more complex due to the increase of functions and connectivity issues (Den Ouden, 2006).

Increasing customer requirements

As said above, producers have to deal with increasing complexity (Brombacher and De Graef, 2001). At the same time meeting customer requirements has become more difficult as a result of the increasing customer expectations. This is also reflected in the warranty period and warranty coverage (Berden, Brombacher and Sander, 2000). In the past only “technical failures” where covered by the warranties. Technical failures are “situations were the product is not able to meet both the explicit (technical) product specifications and customer requirements” (Brombacher et al., 2005). Reliability of a product could be defined as "the probability that a system will perform its intended function for a specified period of time under a given set of conditions" (Lewis, 1996). The classical definition of reliability is: a product fails when it does not comply with its technical specification.

There is a trend that the warranty period will be extended (Berden et al., 2000). The warranty will also move in the direction of covering non technical failures (Berden et al., 2000), see figure 1.2. Non technical failures are: “Situations where in spite of meeting the explicit product specifications, a customer explicitly complains on the (lack of) functionality of the product” (Brombacher et al., 2005). Nowadays also in this situation a customer can return his product. The classical definition of reliability doesn’t cover the problems companies are confronted with. Therefore the classical definition is extended to: “a product fails when it does not comply with expectations of the user” (Brombacher et al., 2005). There is great pressure on increasing customer satisfaction. However the information about these increased customer requirements is not available.

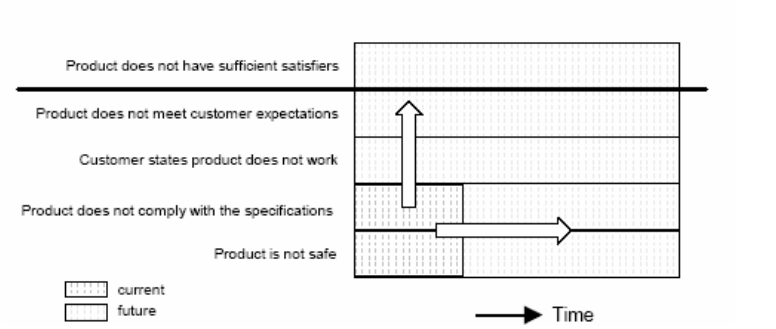


Figure 1.2: Warranty trends in the field of quality and reliability (Berden et al., 2000)

Implications of trends

Manufactures need to develop more complex products, with a shorter time to market and increasing customer demands. Petkova argues that the field feedback information takes too long for timely product quality improvements (Petkova, 2003). Also when ultimately the field feedback reaches the manufacturer, it appears that this feedback information is incomplete and not suitable for performing root cause analysis (Petkova, 2003). Also the increased product complexity makes it almost impossible to identify all the failures in these products and to derive the root cause failure characteristic accordingly (Williams, Banner, Knowles, Dube, Natishan and Pecht, 1998), (Scully, 1998). Due to the unexpected product use and interaction between consumer products and their use environment, replication of the exact failure circumstance might not be possible during failure analysis (Jones and Hayes 2001). Until now current analyses of consumer complaints focus solely on technical failures and not on non technical failures. (Den Ouden, 2006). The current feedback systems are not good enough (anymore) in order to improve the product quality and reliability.

In this situation it is not unexpected that the developed products do not match customer expectations. Due to the incomplete information and information arriving too late manufacturers can not know what their customers expect. The customer requirements as defined by the manufactures and the actual customer expectations could differ (Goodman, 2002). Due to this mismatch a lot of products return to the manufacturer. The manufacturer can not find the failure in the product. The result is a No Failure Found (NFF). A NFF is a complaint of the user where the cause of this complaint could not be found. Conclusively, there is a lack of data in the service centers about non technical failures (Den Ouden, 2006).

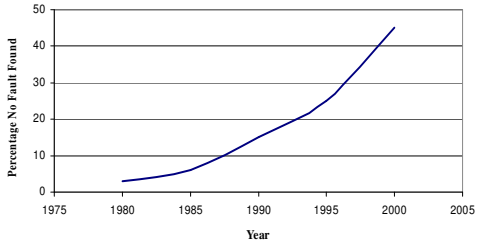


Figure 1.3: Percentages “no failure found” in high-volume consumer electronics (Brombacher, 2002)

There is an enormous increase in the percentage of NFF, see figure 1.3. Products are returned to the manufacturer which are all right according to the manufacturer, but still contain failures in the eyes of customer. Figure 1.3 shows that this trend worsens in time. In other words, critical failures (from a user perspective) are still present in the product at market introduction. The increase in NFF in figure 1.3 suggests that manufactures make the wrong decisions for preventing product failures. Customers and manufactures have a different view on what a product should do and should not do. Due to the increasing customer requirements, manufactures should focus on the specifications and failures considered important

by users (consumers). In other words the user perspective should be taken into account when analyzing product reliability (Den Ouden, 2006).

The current business trends cause an information gap between the developers and users of consumer electronic products (Brombacher et al., 2005). Also there is increased attention for consumer satisfaction combined with a stronger time-to-market pressure. The quality of the feedback from the field does not have the quality for finding root causes of failures and non-technical problems. Due to the lack of this information developers can only use their own technical expertise in failure prioritization. This technical view differs from the user's view. The user perspective is lacking in failure prioritization (De Visser, 2006). Therefore a model should be developed that puts forward the user perspective in the failure prioritization process (Den Ouden, 2006).

These findings give rise to the need for a failure prioritization model from a user perspective. A PhD research is set-up for developing a failure prioritization model from a user perspective. This model should support reliability optimization on two points: Before market introduction, for the prioritization of failures from a user perspective in the development process. After market introduction, for reliability optimization in future product generations. Before market introduction, this model should guide the developers in prioritizing failures. With this model the developers should rate the impact of the different failures of a product the same as users do. After market introduction, the reliability feedback about this product could be analyzed with this model. This results in a failure prioritization which can be used in next generation product improvement. In the PhD research this is translated into the following research question:

How to develop a validated classification model for the determination of the failure severity from a user perspective?

This question was translated in a research model which is given in figure 1.4 below.

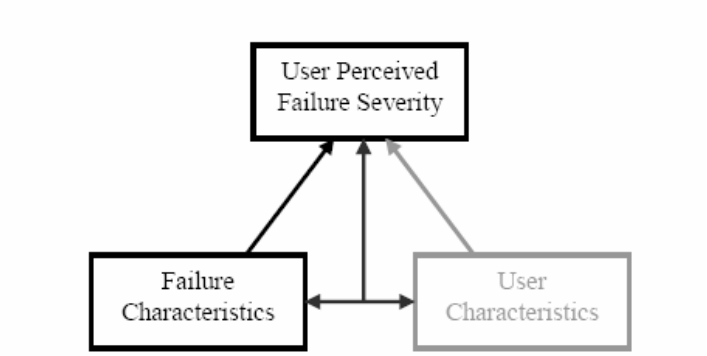


Figure 1.4: High level Research model (De Visser, 2006)

This model has three variables:

- **User Perceived Failure Severity (UPFS):** is the level of user-irritation caused by a product failure.
- **Failure Characteristics**: The attributes of a failure that (possibly) influence the UPFS
- **User Characteristics**: The attributes of a user (group) that (possibly) influence the UPFS.

The preliminary model (figure 1.4) indicates that UPFS is presumably influenced by failure characteristics, user characteristics and the interaction of these two types of characteristics. This model forms the basis of two PhD research projects: one exploring the influence of failure characteristics on UPFS and another exploring the influence of user characteristics on UPFS. The PhD work of De Visser

concentrates on the influence of failure characteristics on UPFS (black parts of the figure) (De Visser, 2006). This master thesis is part of the PhD work of De Visser.

For validating this model, relevant failure characteristics have to be identified. Six characteristics are selected after a thorough investigation of the literature from different scientific fields, and through group discussions with people from industry (involved with product development) and scientists (in quality and reliability) (De Visser, 2006):

- Failure frequency: Number of failures per time unit under standardized use conditions
- Failure impact: The percentage of loss functionality as a results of the failure
- Failure reproducibility: The degree of repeatability of the function affected by the failure
- Function importance: The relative importance of the function affected by the failure
- Failure solvability: The required effort a user should take after failure occurrence to return to normal functioning of the product (excluding the failed part of the function)
- Failure Work Around: The degree in which the failure occurrence can be prevented by the user by operating the product differently

Given the strict time and resource constraints, a complete validation of the hypothetical model within one PhD project is not feasible. Two independent failure characteristics are chosen based on the following criteria:

- Estimated importance of the failure characteristic's relation with UPFS indicated by the participants of the group discussion.
- Practical feasibility of using the characteristic for the failure prioritization process.
- Estimated ease of manipulation in experimental design.

Evaluation of all failure characteristics based on these criteria resulted in the selection of the two following independent variables: failure impact and function importance. The failure prioritization model, the User Perceived Failure Severity (UPFS): defined as the level of user-irritation caused by a product failure central in the PhD work of De Visser is presented in figure 1.5. UPFS is influenced by Failure Impact and Function Importance.

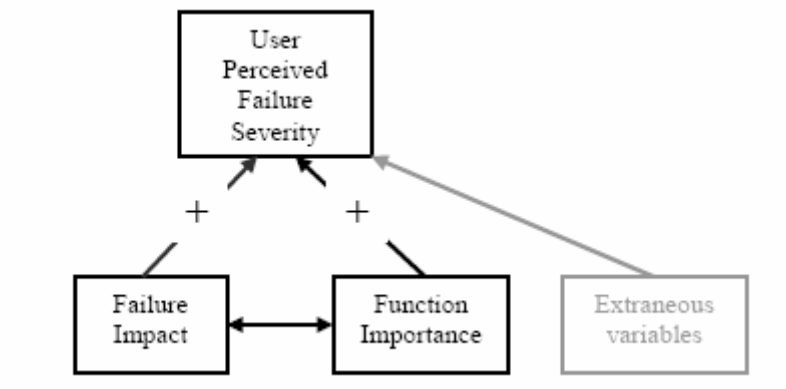


Figure 1.5: User Perceived Failure Severity model

User Perceived Failure Severity (UPFS): is the level of user-irritation caused by a product failure.

Two of the variables expected to influence UPFS are:
Failure Impact: The percentage loss of functionality as a result of the failure
Function Importance: The relative importance of the function affected by the failure

There is also one important user characteristic that may influence the UPFS, the so called "**irritableness"** of people (De Visser, 2006). This user characteristic is an extraneous variable. An extraneous variable is a variable that is not of interest in this study but which might have influence on the relationships being studied (Goodwin, 2005). "Irritableness" can be seen as "how quickly" people get irritated. Some people get irritated by minor product failures other people get irritated by major product failures. Not controlling for the user characteristic "irritableness" makes it unclear if the perceived failure severity is caused by the failure characteristics failure impact and function importance or by the extraneous variable irritableness. This research model is tested with an experiment in a consumer context. Therefore, the following research question has to be answered:

I. ***How to measure and control for "irritableness" of people interacting with a consumer product?***

The impact of the failure and the importance of the failed function cause some irritation at the participants. This user-irritation will be measured with the construct UPFS. UPFS is defined as the level of user-irritation caused by a product failure. So, the second research question which needs to be answered is:

II. ***How to measure User Perceived Failure Severity?***

The four important business trends and their implications show that there is a need for a classification model for the determination of the failure severity from a user perspective. This resulted in a research model consisting out of function importance, failure impact, UPFS and irritableness. Based on that model the objective of this master thesis is twofold: controlling for irritableness and measuring UPFS, see figure 1.6.

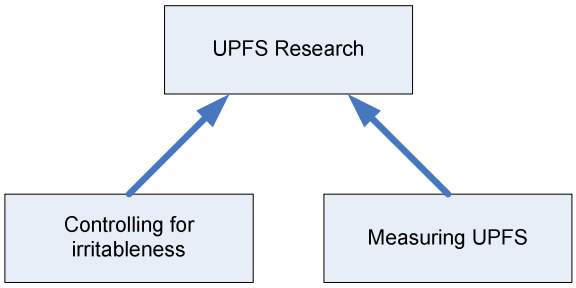


Figure 1.6: Goals Master Thesis project

The next chapter will continue with the methodology for answering these research questions.

2. Methodology

In the previous chapter the research questions have been formulated. In this chapter the approach for answering these research questions is given. These research questions will be answered with the help of an experiment. This experimental approach is explained in subsection 2.2. For helping answering the research questions several research sub questions are formulated in section 2.3. First will be returned to the User Perceived Severity model (section 2.1).

2.1 Research model

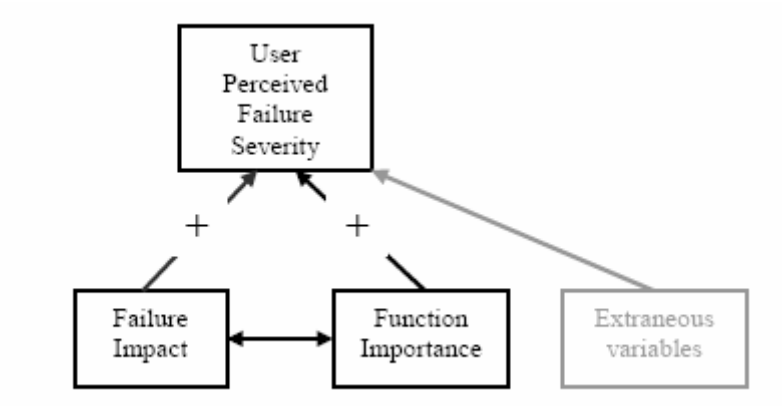


Figure 2.1: Represented User Perceived Failure Severity model (figure 1.5)

Parts of this model presented by De Visser (2006) will be validated by means of three research projects. Firstly, the PhD project of De Visser. Secondly, the Master Thesis project of Gielen. Thirdly, this Master Thesis project of Bergmans. These two master thesis projects contribute to the PhD project of De Visser. The general research question formulated by De Visser needs to be answered with this master thesis project. The general research question is: Increasing the function importance of failures will increase the UPFS significantly. The two specific goals of the master projects are:

- Gielen (2007): To what extent does Function Importance influence UPFS?
- Bergmans (2007): How to control for irritableness? and How to measure User perceived Failure Severity?

All these goals will be tested with one experiment.

2.2 Experimental approach

As said the UPFS-model model (figure 2.1) will be tested with an experiment. When this research project started some preliminary activities with regard to the experiment were already completed. The high level design of the experiment was already finished (see appendix 2A). In short, this means that an experiment will be performed where the user perceived failure severity at two different function importance levels will be measured. These two levels are a low function importance level and a high function importance level. What a low importance and a high importance function is, is examined by Gielen (2007) and will be presented in chapter 7. The participants of the experiment have to interact with a consumer electronic product, more specifically a high end television. Gielen rated these functions of this television on importance from a user perspective. For the experiment participants have to be selected and as formulated in research question I: there has to be controlled for irritableness. During the experiment UPFS has to be measured. This target is formulated in research question II.

This experiment will be conducted in a laboratory, for greater control. Conditions of the study can be specified more clearly, and participants can be selected and placed in conditions more systematically (Goodwin, 2005). Field research matches the settings in daily live more closely. However, laboratory research contributed a lot to the knowledge about behavior. Goodwin (2005) argues that the creation of similarity to daily living is certainly not the most important condition in research. In a study conducted by Anderson, Lindsay and Bushman (1999), a comparison of 288 laboratory and field studies that investigated the same variables, it is concluded that there was a high degree of correspondence between the results in- and outside the laboratory.

A disproportionate distribution of irritableness levels over the experimental groups may reduce the likelihood of finding significant correlations between the independent variables (function importance and failure impact) and the dependent variable UPFS. "Irritableness" is kept constant in the experiment by matching. In matching, participants are grouped together on some specific variable (such as irritableness) and distributed randomly to different groups in the experiment. It is important to keep the groups of participants equal among the experiments, because the measurement of the impact of function importance and failure impact on UPFS is the objective. When the user characteristic irritableness is not kept constant, it is not clear if the results of the experiment are caused by the failure characteristics (function importance and failure impact) or by the extraneous variable irritableness. Therefore, the goal of this master thesis project, is to select and test if instruments can control for irritableness. The second goal is to select an instrument which could measure UPFS. This is translated in the two research questions presented in chapter 1. These research questions will be answered with the help of several research sub questions presented in the next section.

2.3 Research sub questions

Both concepts, "irritableness" and UPFS, involve irritation. Irritableness is defined as how quick people get irritated and UPFS is defined as the level of user-irritation caused by a product failure. Irritation is an emotion, (Wranik, 2005). That is why the following research sub question is formulated:

A. *What is an emotion?*

The answer to this question (A) will give more insight about emotions and specifically the irritation emotion. This gives a better understanding of the concepts irritableness and UPFS. Based on the answer of question (A) it is clear what kind of irritation-emotions are of interest in this research. The literature about measurement scales for (irritation)-emotions will be investigated. With these measurement scales irritableness and UPFS should be measured. This is formulated in the following research sub question:

B. *What measurement scales can measure (the irritation) emotion?*

Emotions can be measured in several ways. Therefore an overview will be given of the possibilities for measuring the (irritation)-emotion(s). Chapter 3 will start with an introduction about emotions (research sub question A) and will continue with how emotions can be measured (research sub question B).

The following two sub-questions (C and D) presented next, are related to the first research question (I):

How to measure and control for "irritableness" of people interacting with a consumer product?

Question B results in some measurement scales which possibly can measure the irritation emotion. A choice has to be made for selecting a scale for the measurement of irritableness. This selection will be

based on criteria. These criteria will be elicited and subsequently a choice will be made. This is formulated in the following research sub question:

C. What are the criteria for selecting a measurement scale for irritableness?

The selected measurement scale needs to be prepared so that it will be applicable in this experiment. A description of the experiment will be given and the application of the selected measurement scale will be outlined. The results on this measurement scale have to be interpreted in such a way that there can be controlled for irritableness. This is formulated in the following research sub question:

D. How will the selected measurement scale be used in the experiment to control for irritableness?

In the first section of chapter 4 selection criteria for a scale measuring irritableness, will be presented (research sub question C). Subsequently, these criteria will be applied to the different instruments and an instrument will be selected for measuring irritableness and UPFS. In the second section will be presented how the instrument will be used in the experiment (research sub question D).

The following two sub-questions (E and F) presented next, are related to the second research question (II):

How to measure User Perceived Failure Severity?

Question B results in some measurement scales which possibly can measure the irritation emotion. A choice has to be made for selecting a scale for the measurement of UPFS. This selection will be based on criteria. These criteria will be elicited and subsequently a choice will be made. This is formulated in the following research sub question:

E. What are the criteria for selecting a measurement scale for UPFS?

The selected measurement has to be adapted that it will be applicable in the experiment. A description of the experiment will be given and the application of the selected measurement scale will be outlined. The results on this measurement scale have to be interpreted in such a way that UPFS can be measured. This is formulated in the following research sub question:

F. How will the selected measurement scale be used in the experiment for measuring UPFS?

In Chapter 5, criteria for selecting a scale measuring UPFS, will be presented (research sub question E) and an instrument will be selected. How this particular instrument will be used in the experiment (research sub question F) is presented in the last section of chapter 5. The selected measurement scales will be evaluated on reliability and validity in chapter 6. In chapter 7 a detailed design of the experiment will be given. This design will be tested with a pilot and based on these results the definitive design of the experiment will be presented. In chapter 8 the results of the experiment will be presented.

3. Measuring emotions

In this chapter first an introduction will be given about emotions. Irritableness and UPFS are related to the irritation emotion. Therefore the irritation emotion will be explained into more detail (research sub question A). Subsequently several methods for measuring irritableness and UPFS will be presented (research sub question B).

3.1 Emotions

In this section the irritation emotion will be explained in more detail. In literature there is no consensus about the question what exactly is an emotion (Wranik, 2005). Scherer views emotions as "the interface between an organism and its environment mediating between constantly changing situations and events and the individual's behavior responses" (Scherer, 1984). Irritation is part of the "anger family" emotion (anger, irritation, contempt and frustration) (Wranik, 2005). Irritation can be considered as a weaker form of anger.

Personality research brings together contributions from developmental, social, cognitive and biological psychology. Personality research studies the whole person, the dynamics of human affect and motivation, and the identification and empirical measurement of the individual differences among persons (McAdams, 1997). Emotions and other affective phenomena are important for human behavior (Wranik, 2005). Affect plays a central role in personality theories. Davitz made a division for person characteristics related to affect: first, stable traits that describe how people usually or typically are; second, transitory states that describe the affective condition of people during a specific moment in time (Davidz, 1969). Emotion researchers are primarily interested in the affective states, whereas personality researchers are primarily interested in affective traits. However, the boundaries between affect-related traits and affect states are vague (Chaplin, John and Goldberg, 1988), (Endler and Magnusson, 1976), (Spielberger, 1972). Both states are part of emotion reaction (Wranik, 2005). Emotions are complex phenomena; "for each level or facet of emotions, different theoretical approaches can be used to understand the relationship between personality traits and specific affective feeling states and responses" (Wranik, 2005).

In the classical paper of the founder of traits theory, Gordon Allport, traits are seen as the building blocks of personality, the guidepost for action, the source of individual's uniqueness (Allport, 1937). Traits are inferred predispositions that direct the behavior of an individual in consistent and characteristic ways (Gibson, Ivancevich, Donelly and Konopaske, 2003). Traits produce consistencies in behavior because they are enduring attributes, and they are general or broad in scope (Gibson et al., 2003). "The fundamental goal of trait theory is to characterize individuals in terms of a comprehensive but finite, preferably small set of stable dispositions that remain invariant across situations and that are distinctive for the individual, determining a wide range of important behaviors" (Mischel and Shoda, 1998).

As said before, for the experiment it is important to keep the factor user characteristics equal among the groups. This will be done by matching. In advance of the experiment irritableness of the participants will be measured. Irritableness of people is related to stable traits. Traits describe how people usually or typically are. Based on this irritableness measurement, the matching procedure will divide the participants in such a way that equal groups are created. There can be controlled for extraneous user effects. For UPFS it is important that the difference in affective states can be measured. A product failure occurs and the reaction of a customer on this failure has to be measured. Therefore for UPFS, measuring the change in the transitory state of consumers is required.

In this section a lot of terms related to emotion are used. The concepts of emotion, affect and mood are used interchangeably by researchers. However, according to Oliver (1997) these concepts can be differentiated. The terms, affect, emotion, mood, trait and state will be defined below. In appendix 3A the

definitions of the relevant terms are given. However, in this research emotions are considered to consist of stable (affective) traits and fluctuating (affective) states see also appendix 3A.

3.2 Measurement scales for emotions

From the previous section it becomes clear that there are stable traits and transitory states. The extraneous variable irritableness, is related to stable traits. User Perceived Failure Severity is related to a change in the affective states caused by a product failure. This product failure causes a change in the transitory state. So, for measuring UPFS the change in transitory states has to be measured. For irritableness the stable traits have to be measured.

There are several measurement techniques for emotions. Four ways of measuring emotions will be given and discussed. The first one is Positive and Negative Affect Schedule (PANAS). The second one is the critical incident question. The third one is related to measuring irritation in commercials. The last one is related to measuring irritation in a working environment. For every method will be explained whether they measure states or traits.

3.2.1 Positive and Negative Affect Schedule (PANAS)

The Positive and Negative Affect Schedule (PANAS) is suitable for measuring traits and (changes) in states. In a study conducted by Ihrig, Hoffmann and Triebig participants were exposed to ammonia vapors during five consecutive days (Ihrig, Hofmann and Triebig, 2006). One of their research aims was to examine the impact of personal traits on the intensity of self-reported health. Their personality traits were measured with the Positive and Negative Affect Schedule (**PANAS**) (See appendix 3B). Participants with high positive affect reported less complaints, then persons with low negative affect. See Appendix 3C for a detailed discussion about PANAS.

The PANAS-X scale is simple and easy to use. Most subjects complete the entire 60-item schedule in 10 minutes or less. Due to time constraints it is also possible to select those scales that are most relevant within the context of the research (Watson and Clark, 1999).

3.2.2 Critical Incident Question

With the critical incident question it is possible to measure emotional states. The research of Van Dolen, Lemmink, Mattsson and Rhoen (2001) explores the effect of emotion on satisfaction with after sales services. Van Dolen et al. (2001) tried a novel approach to examine emotions in service contexts by extracting the emotional content of written answers to the **critical incident question** and relating it to satisfaction, as well as, to the source of emotion. See Appendix 3D for a detailed discussion about the critical incident question.

3.2.3 Irritation in commercials

With the proposed scale for measuring irritation in commercials, the (emotional) state will be measured. De Pelsmacker and Van den Bergh (1998) researched the irritation caused by 226 Belgian commercials, broadcast in 1995. The relationship between ad characteristics and the level of irritation was studied. In this research 104 consumers participated. The respondents had to fill out an irritation scale for each advertisement separately. The irritation scale was a 7-item 7 category Likert scale, developed on the basis of preliminary exploratory research. A Likert scale contains a statement (item) where the respondent is asked to indicate his or her degree of agreement with that statement. The irritation scale items were: ridiculous, stupid, irritating, annoying, exaggerated, “would like to zap”, “gets on my nerves”. The reliability analysis showed a Cronbach’s alpha of 0.95, which is very good. The irritation scale used by

De Pelsmacker and Van den Bergh is reliable. See Appendix 3E for a detailed discussion about irritation in commercials.

3.2.4 Irritation in working environment

The irritation scale proposed in this section measures the emotional state irritation. Dormann and Zapf (2002) investigate the relationship between social stressors, irritation and depressive symptoms. Social stressors consist of social animosities, conflict with co-workers and supervisors, unfair behavior, and a negative group climate (Dormann and Zapf, 2002). In their study the following hypothesis was tested: Irritation mediates the effect of social stressors on depressive symptoms. The respondents were 313 residents of Dresden, Germany. This theoretical model was developed by Mohr (1986). Mohr's model suggests that stressors at work are the starting point of a temporarily ordered sequence of different stress reactions which are causally connected. First irritation emerges, which leads to a decrease in self-esteem and an increase in anxiety. Anxiety leads to depressive symptoms and further reduces self-esteem, which also leads to an increase in depressive symptoms (Dormann and Zapf, 2002). In this study irritation is measured by a 3-item scale developed by Mohr (Mohr, 1986): If other people talk to me, I often react grumpily, I am readily annoyed, react irritated, even when I do not want to do so. See Appendix 3F for a detailed discussion about irritation in working environment.

This scale is updated by Mohr, Rigotti and Müller (2005) to an eight item scale (See Appendix 3G). Their irritation scale describes subjectively perceived emotional and cognitive strain in context of working environment. Irritation is defined as a product of an interaction process between person and environment and not as trait (Mohr, Müller and Rigotti, 2005). This scale measures the behavior of, not switching off from you work (Cognitive Irritation (CI), item 1,2 and 4 Appendix 3G) and also agitated irritation (Emotional Irritation (EI), item 3,5,6,7,8 Appendix 3G). The irritation-scale should only be used for adults with a job or a job history, due to the explicit mention of work related scale-items (Mohr et al., 2005). In their study scale and item characteristics of the irritation scale are presented on the basis of 15 studies (N = 4030).

In this chapter several instruments are proposed for measuring irritableness and UPFS. PANAS-X, critical incident question, irritation scales used for commercials and the irritation scale of Mohr et al can measure (affective) traits. PANAS-X and the critical incident question also can measure differences in (affective) states. In the next chapter selection criteria will be presented for the measurement scales of irritableness and UPFS. Subsequently, based on these criteria, a decision will be made for how to measure irritableness and UPFS.

4. Measuring irritableness

In the previous chapter some methods are presented for measuring emotions: stable (affective) traits and transitory (affective) states. In this chapter will be assessed if these methods could measure irritableness, a stable (affective) trait (figure 4.1). Therefore the suitability of these methods for measuring irritableness will be compared in this chapter. This comparison is based on criteria which will be given below (research sub question C). Subsequently the selected instrument will be explained in more detail with attention to how this measurement instrument can measure irritableness (research sub question D).

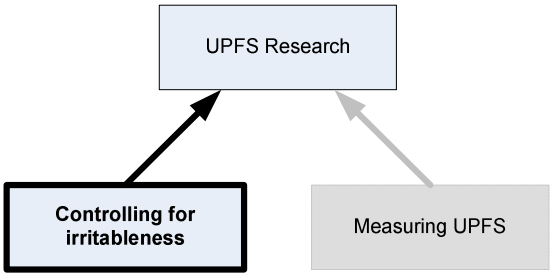


Figure 4.1: Goal chapter 4

4.1 Assessment of irritableness measurements

In the previous chapter some methods are presented for measuring emotions. Irritableness is derived from these emotions. In this section the methods for measuring emotions will be compared. This comparison is based on criteria which are presented below. Based on this comparison a measurement instrument will be selected for measuring irritableness.

The options given in the previous chapter for measuring irritableness will be evaluated. Hughes (1967) defined three criteria for determining a scale's (c.q. measurement's) usefulness:

1. Does the measurement instrument measure the variable of interest?
2. Is this measurement instrument reliable (and valid)?
3. Is it easy (enough) to use?

For the first criteria the measurement scale has to measure irritableness. Irritableness can be seen as "how quickly" people get irritated. Some people get irritated by minor product failures other people get irritated by major product failures. Also, the measurement has to be constructed in such a way that it does not change a person state. Some measurements scale's can only be used when the participant has to react on an incident. This incident can be evaluated positively or negatively. When using the results obtained by the "irritableness" measurement scale in this situation, participants are extremely in a positive mood or in a negative mood. For the experiment, the objective is to measure how irritated the participants are in general. So, the personal traits, which are stable over time, are the matter of interest here. Therefore the measurement scale has to be applicable in a normal situation which is not influenced on purpose with a particular incident.

For the second criteria, the reliabilities of the measurement scales will be compared based on their Cronbach's alpha, a measure that assesses the consistency of the entire scale (Hair et al., 2005). . Cronbach's alpha is a measure that assesses the consistency of the entire scale. In one case, critical incident analysis, Cronbach's alpha is not mentioned, an alternative measure, the interrater reliability is used. In this case, this is a measure of agreement between two judges on the assessment of the incidents.

For the third criteria the easiness has to be such that participants who are exposed to this measurement understand the meaning of this measurement. Also the time the participant is occupied with the measurement of irritableness has to be short. The participants participate voluntary in this experiment. So, a measurement scale which takes too much time of the participant and is difficult to work with will lead to a small group of persons willing to participate in this experiment. Also the participants will lose their attention and the "quality" of the results obtained by this measurement scale will decrease. The measurement possibilities mentioned in the previous section will be evaluated against these three criteria for this research below:

PANAS-X

Criteria 1: PANAS-X measures, among other things, Negative Affect. Negative Affect measures irritation in an extensive way. Irritableness can be expressed in a lot of ways. Some one can feel sad, another person becomes angry. The way of expressing emotions is person specific, i.e. depending on someone's personal traits. With PANAS-X it is possible to capture all these expressions of irritableness. So, the PANAS-X measures irritableness as intended by this research.
Criteria 2: Reliability and validity are good (with one exception for the subscale Surprise). This subscale surprise is not relevant for this research because irritableness is not related to surprise. The alpha reliabilities are ranging from .83 to .90 for the general Positive Affect, and from .85 to .90 for general Negative Affect (Watson and Clark, 1999). A drawback is that only a part of PANAS-X, the General Dimensions Scales, has been translated in Dutch and validated by Hill et al. (2005).
Criteria 3: The PANAS-X is easy applicable in the form of a questionnaire. The entire 60-item schedule can be completed in 10 minutes or less (Watson and Clark, 1999). If necessary this time can be reduced by only selecting those scales that are most relevant to this research.

Critical Incident Analysis

Criteria 1: Since the respondents are exposed to a critical incident, the participants are not in a "neutral" state anymore. This will bias the results with regard to the general "irritableness" of people. So, this measurement does not measure irritableness as intended by this research.
Criteria 2: The reliability is sufficient. Interrater reliability ranged from 86% to 100%. However, it has to be noticed that their "classification system could be considered as the starting point in determining the proper categorization of emotions evoked by critical incidents" (van Dolen at al., 2001). Probably some emotions are not captured in the classification system and other emotion categories are not relevant (van Dolen at al., 2001).
Criteria 3: A suitable critical incident has to be developed for this research. However, developing a suitable critical incident conflicts with criteria 1.

Irritation commercials' scales

Criteria 1: Watching commercials will influence the mood of the participants. So, the participants react on an intervention and lose their "neutral" state. This will bias the results with regard to general "irritableness". So, this measurement does not measure irritableness as intended by this research.
Criteria 2: Reliability is sufficient. Cronbach's alpha ranged from 0.87 to 0.95 .
Criteria 3: Some of the items used are related to commercials: "would like to zap" and "the commercials constituted a pleasant break from the program". These items are not applicable when the scales are not adapted for this research project.

Irritation scale of Mohr, Rigotti and Müller

Criteria 1: The irritation scale of Mohr, Rogotti and Müller (2005) does not measure a trait or state, it measures work related stress. Since irritableness is related to personal traits this measurement scale is not applicable in this research.

Criteria 2: The reliability is sufficient. Cronbach's alpha ranges from 0.84 to 0.93. However the validity is questionable: some items correlate high on both factors Cognitive Irritation and Emotional Irritation.
Criteria 3: The irritation scale of Mohr, Rigotti and Müller (2005) is easy applicable in the form of a questionnaire.

Conclusively this evaluation can be summarized in the table presented below.

Table 4.1: Comparison of measurement instruments for irritableness

Measurements	**Criteria**		
	1. Measures irritableness?	2. Reliability and (Validity)?	3. Easiness?
PANAS X scale	++	++	++
Critical Incident Analysis	-	+	_
Irritation commercials' scales	-	+	_
Irritation scale	-	+/-	++

Note: ++ : very good; + : good; +/-: questionable; - : not sufficient

Based on the analysis above, measuring irritableness as a stable trait can best be done by use of PANAS-X. PANAS-X has been used widely, and is proven to be reliable, valid and easy to use.

4.2 Use of PANAS-X for measuring irritableness

In the previous section it became apparent that using PANAS-X for measuring irritableness is the best option. In this section it will be explained in more detail how PANAS-X will be applied in this master thesis project.

As said before a disproportionate distribution of "irritableness" levels over the experimental groups will bias the results. The likelihood of finding significant correlation between the independent variable function importance and the dependent variable User Perceived Failure Severity will be reduced. By using PANAS-X the irritableness of participants can be measured. Since the objective of the experiment is to measure the impact of different failure characteristics on UPFS, the participant's characteristics between the groups have to be as equal as possible. So the participants "irritableness" level has to be average and kept equal among the groups. PANAS-X consists of several scales (Appendix D). The usefulness of these scales for measuring irritableness, will be discussed below.

In a study conducted by Ihrig, Hoffmann and Triebig people were exposed to several levels of ammonia. Consequently complaints about irritative symptoms, respiratory symptoms and olfactory symptoms were reported (see Appendix 4A) for definitions of these symptoms) (Ihrig et al., 2006). The goal of their study was to examine the impact of personal traits and regular contact at the workplace on the intensity of self-reported health symptoms and complaints.

The researchers used the original 20-item PANAS for measuring personal traits. The original PANAS has two 10-item dimensions: Positive Affect and Negative Affect. People with high positive affect reported less irritative symptoms, where as people with negative affect report more olfactory and respiratory symptoms (Ihrig et al., 2006). It has to be noted that these findings decrease with higher exposure levels of ammonia. So, based on this research the scale Negative Affect is an important scale for measuring "irritableness" of people. People with high Negative Affect reported more complaints so; these people get more quickly irritated than other people. Consequently the Scale Negative Affect should be selected for

measuring “irritableness”. The General Dimension Scale Negative Affect will be used for measuring irritableness. The Basic Negative Emotion scales Fear, Hostility, Guilt and Sadness of PANAS-X (Appendix 3C) are consistently and substantially intercorrelated (loadings range from 0.69 to 0.83) with the General Dimension Scale Negative Affect. Thus, it could be expected that a high score on the subscales of the Basic Negative Emotion scales will lead to a high score on the General Dimension Scale Negative Affect. Therefore, also these Basic Negative Emotion scales will be used for measuring irritableness. The Other Affective Scales Shyness, Fatigue, Surprise and Serenity do not load strongly or consistently on one of the two scales (Watson and Clark, 1999). However, only using the scales related to Negative Affect will lead to response acquiescence – a tendency to agree with statements (Goodwin, 2005). In this case this would result in a case that participants have a too high negative affect. To avoid these problems, surveys with Likert scales typically should balance favorable and unfavorable statements (Goodwin, 2005). For the experiment this means that not only Negative Affect Scales should be incorporated but also Positive Affect Scales and Other Affective Scales should be incorporated. This should force participants to read each item carefully and make item-by-item decisions (Patten, 1998), from which the quality of the measurement will benefit.

5. Measuring UPFS

In Chapter 3 it became clear that two measurement instruments; PANAS-X and Critical Incident Analysis, could measure transitory (affective) states. The goal of this chapter is to compare these two methods and choosing the best method for measuring UPFS (figure 5.1). This comparison is based on criteria which will be presented below (research sub question E). Based on this comparison a measurement instrument will be chosen for measuring UPFS in the experiment (research sub question F).

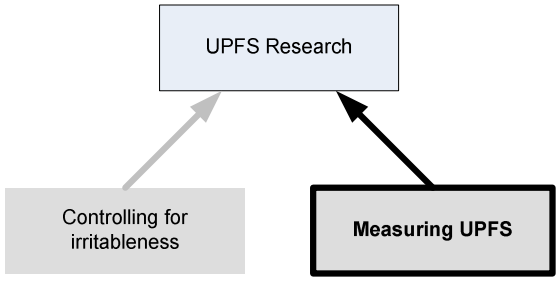


Figure 5.1: Goal chapter 5

5.1 Assessment of UPFS measurement

First a brief summary of the main findings of section 3.2 are given. In the experiment the dependent variable is UPFS. UPFS is the level of user-irritation caused by a product failure. During the experiment a failure occurs. The reaction on this failure will be measured with the UPFS measurement tool. The failure that happens during the experiment can be seen as a critical incident. So, the critical incident analysis might be a good measurement scale. Since, the PANAS-X also measures affective states and differences between states (Watson and Clark, 1999) this measure also can be used for measuring UPFS. The difference between the scores on the PANAS-X scale at the start of the experiment (PANAS-X-1) and at the end of the experiment (PANAS-X-2) can be seen as a measure of UPFS (see figure 5.2). At the start of the experiment, participants have a certain irritation level. The failure which occurs during the experiment will enlarge (or not enlarge) the irritation. This possible enlargement of the level of irritation can be used as a measurement of the UPFS. These two options will be evaluated more structured below.

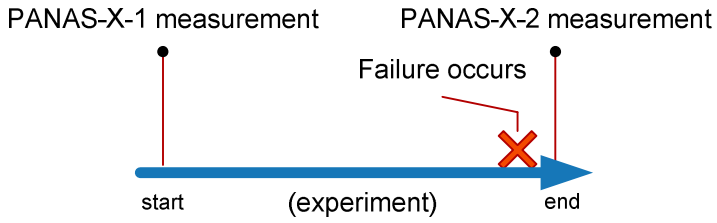


Figure 5.2: Use of PANAS-X for measuring UPFS

The options given above for measuring UPFS will be evaluated by three criteria for determining a scale's (c.q. measurement's) usefulness (Hughes, 1967):

1. Does the measurement instrument measure the variable of interest?
2. Is this measurement instrument reliable (and valid)?

3. Is it easy (enough) to use?

For the first criteria the measurement scale has to measure UPFS. User perceived Failure Severity is the level of user-irritation caused by a product failure. The occurrence of a failure is a short during event which changes the mood of the participants. The objective is to measure this change in mood caused by a product failure. In comparison with measuring irritableness, the objective of UPFS is not to measure how irritated persons are in general. Instead the mood fluctuations are the matter of interest, with other words the change in emotional states.

For the second criteria, the reliabilities of the measurement scales will be compared based on their Cronbach's alpha. Cronbach's alpha is a measure that assesses the consistency of the entire scale. In one case, critical incident analysis, Cronbach's alpha is not mentioned, an alternative measure, the interrater reliability is used. In this case, this is a measure of agreement between two judges on the assessment of the incidents.

For the third criteria the easiness has to be such that participants who are exposed to this measurement understand the meaning of this measurement. Also the time the participant is occupied with the measurement of irritableness has to be short. The participants participate voluntary in this experiment. So, a measurement scale which takes too much time of the participant and is difficult to work with will lead to a small group of persons willing to participate in this experiment. Also the participants will lose their attention and the "quality" of the results obtained by this measurement scale will decrease. This is especially important for measuring UPFS. When a measurement takes too long, participants will return to their normal trait level of irritation. When this is the case, changes in mood fluctuations will not be measured and subsequently criteria 1 will not be satisfied. Another point of easiness to use is the easiness for the researcher. This point is not that much related to the time it takes to analyze the results but merely to knowledge needed to interpret the results obtained with this scale. Some methods expect from the researchers a high level of expertise on emotion research, like the Critical Incident Analysis. Currently this level of expertise is not available within this project.

The two measurement possibilities, PANAS-X and Critical Incident Analysis, will be evaluated against these three criteria.

PANAS-X

Criteria 1: PANAS-X measures Positive Affect and Negative Affect. Negative Affect measures irritation in an extensive way. This irritation can be expressed in a lot of ways. Some one can feel sad, another person becomes angry. With PANAS-X it is possible to capture all these expressions of UPFS. So, the PANAS-X measures UPFS as intended by this research.
Criteria 2: Reliability and validity (discriminant and convergent validity) are good (with one exception for the subscale Surprise). This subscale surprise is not relevant for this research because UPFS is not related to surprise. The alpha reliabilities are ranging from .83 to .90 for the general Positive Affect, and from .85 to .90 for general Negative Affect (Watson and Clark, 1999). A drawback is that only a part of PANAS-X, the General Dimensions Scales, has been translated and validated by Hill et al. 2005.
Criteria 3: The PANAS-X is easy applicable in the form of a questionnaire. The entire 60-item schedule can be completed in 10 minutes or less (Watson and Clark, 1999). If necessary this time can be reduced by only selecting those scales that are most relevant to this research.

Critical Incident Analysis

Criteria 1: The respondents have to react at a critical incident. In this case the critical incident will be the function failure of a television. The reaction is written response. Based on this response the severity of the failure can be measured. So, this measurement does measure UPFS as intended by this research.

Criteria 2: The reliability is sufficient. Interrelated reliability ranged from 86% to 100%. However, it has to be noticed that their "classification system could be considered as the starting point in determining the proper categorization of emotions evoked by critical incidents" (van Dolen at al., 2001). Probably some emotions are not captured in the classification system and other emotion categories are not relevant (van Dolen at al., 2001).
Criteria 3: Participants have to give a written reaction. Framing thoughts about the function failure into words could be difficult for participants. Also for the researchers it is more extensive and difficult to analyze these written reactions.

Conclusively this can be summarized in the table presented below.

Table 5.1: Comparison of measurement instruments for UPFS

Measurements	**Criteria**		
	1. Measures irritableness?	2. Reliability and (Validity)?	3. Easiness?
PANAS X scale	++	++	++
Critical Incident Analysis	+	+	+/-

Note: ++ : very good; + : good; +/-: questionable; - : not sufficient

So, based on the criteria above, PANAS-X is most suitable for measuring UPFS.

5.2 Use of PANAS-X for measuring UPFS

A failure will be perceived negatively. Persons who are externally oriented will look for the causality of failure externally and therefore have more anger, irritation and frustration. Internal oriented persons search for failure causality internally and experience more embarrassment, shame, and guilt (Lazarus, 1991), (Scherer, 2001), (Weiner, 1986). Anger, irritation, shame and guilt are all items of the General Dimension Scale Negative Affect. So, the General Dimension Scale Negative Affect can be used for measuring UPFS.

The PANAS-X scale also has four negative basic emotions scales: Fear, Hostility, Guilt and Sadness. The irritation emotion is part of the "anger family" consisting of anger, irritation, contempt and frustration (Wranik, 2005). "The typical instigation to anger is a value judgment. More than anything else, anger is an attribution of blame" (Averill, 1983). Anger is directed to others. "Anger involves an attribution of responsibility, an accusation, that the target has done something wrong. It follows that the target of anger must be a person or object… to whom responsibility can be assigned (Averill, 1982)". In the experiment the anger of the participants will be directed to the television. The television will "make" the failure. The television will be blamed. Anger will occur in negative, unexpected, and important failure situations, when causality is directed external (Weiner, 1986). During the experiment an unexpected failure will happen. As said before this failure might be an important failure. The basic negative emotion scale Hostility of PANAS-X reflects the emotion anger very well. Firstly, angry and irritation are items of the scale Hostility. Secondly the items hostile, scornful, disgusted and loathing are emotions which are pointed to external factors. These items emphasize that anger is directed to externals. The Basic Negative Emotion Scale Hostility can be used for measuring UPFS.

The basic negative emotion scale Guilt is related to internal causal attribution. Guilt is associated with blaming oneself (Smith and Lazarus, 1993). Guilt is a negatively valanced emotion that typically arises in response to personal failure. (Mascolo and Fischer, 1995), (Roseman, 2001), (Scherer, 2001), (Weiner, 1986). Guilt can arise in this experiment. There are participants who blame themselves for the not correctly functioning of the television (Internals). Internals see the control of their lives as coming from

inside themselves (Gibson et al., 2003). When things go wrong they blame themselves. In an early performed experiment in the Eindhoven University of Technology people were confronted with preprogrammed failures. In some cases people blame themselves when a failure occurred within teletext. The subscale Guilt posses among other things the items: angry a self, disgusted with self and dissatisfied with self. These items clearly measure the self-blame. The subscale Guilt will be used in the measurement of UPFS.

Fear can be associated with evaluating one situation as threatening (Smith and Lazarus, 1993). Weiner, Russel and Lerman concluded that in failure situations, the most reported emotions were anger, depression, fear and frustration (Weiner, Russel and Lerman, 1979). So, the severity of a failure impact can also be expressed in fear. The basic Negative Emotion Scale Fear will also be included in the measurement of UPFS.

Sadness is associated with helplessness in an undesirable situation where there is little or no hope of improvement (Smith and Lazarus, 1993). Persons who internally oriented look for causality of a failure internally and so experience more embarrassment, shame and guilt (Lazarus, 1991), (Scherer, 2001), (Weiner, 1986). People who search for internal causes for negative events (like a function failure of a television) will do more self-blaming and have “more negative emotions in general or more depression/ sadness” (Wranik, 2005). So, also the subscale Sadness will be used for measuring UPFS. Since the basic emotion scales Fear, Hostility, Guilt and Sadness form the Basic Negative Emotion scale, also the Basic Negative Emotion Scale should be included in the measurement of UPFS. Also in this case, only using the scales related to Negative Affect will lead to response acquiescence. So, for the measurement of UPFS also the Positive Affect Scales and Other Affective Scales should be incorporated.

6. Validating PANAS-X

In chapter 4 and 5 is concluded that PANAS-X is suitable for controlling irritableness and for measuring UPFS. In this chapter the reliability and validity of PANAS-X for measuring UPFS and controlling irritableness will be assessed. First a short description will be given about how PANAS-X is translated into Dutch. This Dutch version of PANAS-X will be validated with the help of a survey. In the second section this survey will be explained in more detail. In the third section the reliability and validity of the PANAS-X scales is examined.

6.1 Dutch PANAS-X

The first purpose of using PANAS-X is to select participants who have an average and equal irritableness level who can participate in the experiment. A questionnaire was developed for approaching students. This questionnaire was developed in cooperation with two other persons: the master student who investigates the relationship between function importance and UPFS, and the PhD student who develops a validated classification model for the determination of the failure severity from a user perspective.

A survey was conducted at the Eindhoven University of Technology in the Netherlands. Students of the university in Eindhoven filled in this survey. On objective of this survey was to select a group of persons for the experiment which is as homogeneous as possible. To have a group as homogeneous as possible, only Dutch students were selected. It could be expected that students are a homogenous respondent group in consumer research (Peterson, 2001). So, PANAS-X was translated into Dutch. The general dimension scales were already translated and validated (Hill, van Boxtel, Ponds, Houx and Jolles, 2005). The basic emotion scales and the other affective states were not translated. Two graduate students who speak fluent English and Dutch translated these scales of PANAS-X independently. A committee with the researcher of this thesis and the two graduate students examined the preliminary version. This process was done on consensual agreement. This resulted in the Dutch version of PANAS-X (see also appendix 6A).

6.2 Pre-experimental phase

As said above in chapter 4 and 5 is concluded that PANAS-X is suitable for measuring irritableness and UPFS. However validation is necessary before PANAS-X can be used. The Dutch version of PANAS-X presented in the previous paragraph needs to be validated since this Dutch version has not yet proven to be reliable and valid. The way how PANAS-X can be used and will be validated given in appendix 6B.

In the following section the 200 completed PANAS-X questionnaires will be used for assessing reliability and validity of PANAS-X.

6.3 Reliability and Validity of PANAS-X

The collected data as described in the previous section will be used for assessing the reliability and validity of PANAS-X. The Dutch version of PANAS-X is specially developed for this master thesis project, the PhD project of De Visser (2006) and the master thesis project of Gielen (Gielen, 2007). Therefore the validity and reliability has to be proven. In subsection 6.3.1 a data examination is done with regard to missing values. A lot of missing values can affect the generalizability of the results (Hair et al., 2005). In subsection 6.3.2 a comparison will be made between the results of this study and the results of the Maastricht Aging Study (Hill et al., 2005) with regard to PANAS. With this analysis a part of PANAS-X can be validated. The “other” part will be validated with a factor analysis in subsection 6.3.3. In 6.3.3.1, first the assumptions underlying factor analysis are assessed. Secondly in 6.3.3.2, the factor

analysis itself is executed. In subsection 6.3.4 the reliability of the subscales are examined. Lastly, in subsection 6.3.5 the validity of PANAS-X is checked.

6.3.1 Data examination

In this sub section the data (N=200) is examined with regard to missing values. Too much missing data can affect the generalizability of the results (Hair et al., 2005). The sample size available for analysis is reduced and statistical results based on a nonrandom missing data process could be biased (Hair et al., 2005). A descriptive analysis of the data showed that from the 200 cases in 7 cases 1 item was not filled in and that in 1 case 2 items were not filled in. So, in total 9 values of the 200*60 = 12.000 values were missing (< 1%). The missing values are "replaced" with the predicted values and a complete dataset of 200 cases will be used for further analysis. For a detailed discussion see Appendix 6C

6.3.2 Comparison with Maastricht Aging Study

PANAS-X consists of two higher order scales, the so-called General Dimension Scales: Positive Affect and Negative Affect (see also appendix E). These two scales consist each of 10 items. These two scales cover PANAS. The two scales Positive and Negative Affect of PANAS are validated by Hill et al. (2005). The results of this analysis will be compared with the data collected as described in section 5.2. The analysis conducted by Hill et al. (2005) is called "Maastricht Aging Study". The analysis of the data as described in section 5.2 will be called in this section: "Failure Severity Study".

The analysis of the Failure Severity Study is done in exactly the same as in the Maastricht Aging Study.. The two factors explain in both studies nearly the same percentage of variance (around 38%). The internal consistencies of the factors of both studies are around the 0.80 (Cronbach's alpha). However it has to be noted that Cronbach's alpha of the PA factor is lower in the Failure Severity Study (0.75) than the Maastricht Aging Study (0.84). The reliability of PA is lower in the Failure Severity Study, but is still above the minimum recommended level of 0.7. Besides these differences the results are nearly the same for both studies. These results confirm that the 20 items forming PANAS are reliable and can be used for further analysis of irritableness and UPFS. See for a detailed discussion Appendix 6D.

6.3.3 Factor analysis

In addition to the two higher order scales there are also 11 specific affects within PANAS-X (not within PANAS). For a detailed discussion see Appendix 6E. PANAS-X can be described as "a hierarchical taxonomic scheme in which two broad, higher order dimensions are each composed of several correlated, yet ultimately distinguishable states" (Watson and Clark, 1999). With factor analysis it will be assessed whether the proposed structure also exist within the Dutch version of PANAS-X. With this factor analysis the two higher order dimensions, Positive and Negative Affect, are assessed. Also the 11 specific states will be examined. First however, the assumptions underlying factor analysis will be checked. These assumptions are met. For a detailed discussion see Appendix 6F

6.3.3.2 Performing factor analysis

The results of the previous section show that it is useful to do a factor analysis for data summarization. See also Appendix 6G.

A principal component factor analysis with Varimax rotation was done. The results are presented in Appendix 6H. The mean and the standard deviation are given of each item. Also the factor pattern matrix is given. The factor pattern matrix, contains the factor loading of each variable on each factor. The results show that the items which are related to the General Dimension Scale Negative Affect and the Basic Negative Emotion Scale load on one factor. Only the item eenzaam ("lonely") loads also negatively on

the second factor. The items which are related to the General Dimension Scale Positive Affect and the Basic Positive Emotion Scale load on the other factor. One exception is the item "oplettend" (alert), this item does not load on the second factor (and also not on factor one). For the items related to other affective states no clear pattern is visible, as expected. These results show that there are two general dimensions. The items "lui", "kalm" and "ontspannen" (sluggish, calm and relaxed, respectively) do not load on either component 1 or component 2. These three items are part of the scale affective state Fatigue and Serenity. So, based on the above results the items which should measure negative (or positive) affect also measure negative (or positive) affect in this data-set. The "other" affective state Shyness (with the items verlegen beduusd bedeesd and timide) loads on factor 1. The "other" affective state Surprise (with the items verwonderd verbaasd and versteld staan) also merely loads on factor 1. It has to be noted that "verwonderd" (amazed) also load on factor 2. The "other" affective state Fatigue is also merely related to factor 1 except for the item "lui" (sluggish) which does not load on factor 1 or factor 2. So, based on these results it can be concluded that Shyness Fatigue and Surprise are more related to negative affect. With regard to the other affective state Serenity, two of the three items, namely "kalm" and "ontspannen" (calm and relaxed), do not load on factor 1 or factor 2. No conclusion can be drawn about whether this state is related to negative or positive affect. The overall structure of the two broad higher dimensions is confirmed. See also Appendix 6I.

The next step is assessing the lower order scales of the PANAS-X. For assessing the lower order scales a range of factor solutions was examined, until a solution was reached that contained the most interpretable and clear factor structure. This means that each solution contained a different number of extracted factors. A solution with 6 extracted factors gives the best interpretable result. Interpretable means in this context a clear factor solution: an item loads on one factor. This solution and discussion is shown in Appendix 6J. Also for this factor analysis, a principal component factor analysis and a varimax rotation was used. The conclusion will be discussed below.

So, in conclusion the current data set supports the two general affective states of positive and negative affect. The lower order scales of positive affect can be confirmed with this dataset. The underlying structure of positive affect in the form of the basic emotion scales of Joviality, Self-Assurance and Attentiveness is clearly visible. The underlying structure of Negative Affect can not be confirmed as proposed by Watson and Clark. Nevertheless, the dataset clearly confirms the existence of the "other" affective state Fatigue. Cultural differences could be a reason for sticking together of the four affective states Fear, Hostility, Guilt and Sadness in the Dutch version of PANAS-X. Cultural variations could influence the emotion process (Mesquita and Frijda, 1992). There are strong cultural differences in perceiving emotions. Probably Dutch people do not express their negative feelings in a specific way like Fear, Hostility, Guilt and Sadness. They do not distinguish between these specific basic negative emotions, and perceive all theses basic negative emotions the same, i.e. as one negative affect.

6.3.4 Reliability of PANAS-X

The reliability (Cronbach's alpha) of the subscales of the PANAS-X will be discussed below.

Cronbach's alpha for the General Dimension Scale Negative Affect is 0.842. The generally agreed lower limit for Cronbach's alpha is 0.70 (Hair et al., 2005). However this threshold should be raised as the number of items increases, especially as the number of items approaches 10 or more (Hair et al., 2005). The Basic Negative Emotion Scale has a high Cronbach's alpha (0.906). This is not surprisingly since Cronbach's alpha raises when the number of items also increases (Hair et al., 2005). Also the scales Fear Hostility Guilt and Sadness have sufficient Cronbach's alpha's (0.782, 0.771, 0.755 and 0.779 respectively).

The Cronbach's alpha's for the General Dimension Scale Positive Affect is 0.747, indicating that the reliability of the General Dimension Scale Positive Affect is acceptable. The Basic Positive Emotion Scale has a high Cronbach's alpha (0.837). Also here has to be noted that a high number of items will increase Cronbach's alpha. The Basic Positive Emotions subscales range from (0.666 to 0.813) which can be classified as acceptable. The Cronbach's alpha of the scale Joviality and Self-Assurance is sufficient. The Cronbach's alpha of Attentiveness is 0.666. This value is acceptable considering that this scale only exist out of 4 items.

The Cronbach's alpha for the Shyness scale is 0.582. This value is rather low and it is questionable if the scale Shyness is reliable enough. The scale Fatigue and Surprise have a sufficient Cronbach's alpha which shows that these scales are reliable. The scale Serenity has an unacceptable low value of Cronbach's alpha: 0.350. Therefore the scale Serenity is not reliable. Conclusively, it can be said that the scales of PANAS-X are reliable except Serenity and care should be taken when using the scale Shyness.

6.3.5 Validity of PANAS-X

Besides reliability also the validity of PANAS-X has to be tested.

The convergent validity of the basic emotions scales Fear, Hostility, Guilt and Sadness is sufficient. Also the convergent validity of Joviality, Self-Assurance and Attentiveness is sufficient. Since the correlation between the scales Fear, Hostility, Guilt and Sadness is high (0.502 – 0.654) there is no ground for claiming that these scales posses sufficient discriminant validity. The correlation between Joviality, Self-Assurance and Attentiveness is small enough (0.281 – 0.414) to state that the discriminant validity is sufficient. With regard to the "other" affective states Shyness, Fatigue, Serenity and Surprise no clear conclusion can be drawn about convergent validity, because it is unclear what these other affective states should measure. Discriminant validity is not sufficient due to the many correlations of these scales with other scales. In general it is appropriate to use the general dimensions scales negative affect and positive affect, the basic negative and positive emotion scale, and the scales Joviality, Self-Assurance, Attentiveness. Care has to be taken when using Fear, Hostility, Guilt and Sadness because the discriminant validity is questionable. Using the "other" affective states Shyness, Fatigue, Serenity and Surprise is not recommended. Further research related to user perceived failure severity should not use the scales related to the "other" affective states because, no conclusion concerning the validity of these scales can be drawn. See appendix 6K for a detailed discussion.

6.4 Selection of Participants

In the previous section is concluded that (parts of) PANAS-X are reliable and valid enough to use further in this research. One purpose of using PANAS-X is to select participants who have a an average and equal irritableness level.

There are no guidelines in literature for a normal level of irritableness. The average irritableness level of the 200 participants will be seen as normal. Potential normal participants are those participants who have a score on the General Negative Affect Scale of 18.52 (mean) +/- 5.44 (one standard deviation) and who score on the Basic Negative Affect Scale 9.99 (mean) +/- 2.81 (one standard deviation). A frequency plot for General Negative Affect and Basic Negative Affect is given below (see appendix 6L). The thick black lines in the figure are the lower and upper limit of the ranges. The choices for setting these cut-off values are not validated or discussed in literature. However with this policy it will be ensured that participants with extreme low or high negative affect will be excluded from the experiment. As said before a disproportionate distribution of "irritableness" levels over the experimental groups will bias the results. The likelihood of finding significant correlation between the independent variable function importance and the dependent variable User Perceived Failure Severity will then be reduced. By using the General

Negative Affect Scale and Basic Negative Affect Scale irritableness of participants can be measured. Since the objective of the experiment is to measure the impact of different failure characteristics on UPFS, the participant's characteristics between the groups have to be as equal as possible. The participants "irritableness" level has to be kept equal among the groups. Applying these cut-off values resulted in the selection of 71 participants out of 200.

In summary, in this chapter the first three activities, the pre-experimental phase, were completed, see Appendix 6B. Furthermore, PANAS-X was validated, there can be controlled for irritableness and participants were selected. In the next chapter the activities related to the experimental phase will be discussed.

7. Experiment

In this section the experimental phase will be explained in more detail. In section 7.1 the goal of the experiment will be described. In section 7.2 the experimental design will be shown. In section 6.3 this design is tested with a pilot and the results of this pilot are used for redesigning the experiment.

7.1 Goal of Experiment

As said before, this master thesis project is a part of a larger research project, the PhD work of De Visser. The goal of this experiment is testing the hypothesis that increasing function importance will increase the UPFS. This hypothesis is one of the hypothesis formulated by De Visser (2006). Another master student, Gielen investigates the issues with regard to function importance (Gielen, 2007). In the master thesis project of Gielen it is, among other things, examined which functions are seen as important by users and which functions are seen as less important. The goal of this master thesis project, is to test if the proposed tool PANAS-X can control for irritableness. The second goal is to test if the proposed tool PANAS-X could be used to measure UPFS. All goals mentioned above are tested in one experiment which will be described below.

7.2 Design of the Experiment

Based on the results of the questionnaires about PANAS-X and Function Importance a selection of possible participants is made (section 6.4). Gielen had also several selection criteria with regard to Function Importance (Gielen, 2007). The people who meet both criteria were invited for participating in the experiment.

Experiment in a nutshell

A short description of the experiment is given. In the experiment the participants used a Philips LCD-TV. A living room (see Appendix 7A) was created in a laboratory where they had to "watch" television and to use particular functions of this television. Failures were caused in these functions during the experiment. These failures involved high importance functions and low importance functions. Gielen examined which functions were high important and which one were less important (Gielen, 2007). The highest rated function was to watch the desired program, without failures in screen and sound. The function which was used as a low importance function is the use of the motorized swivel. The motorized swivel can be used to turn the screen between + and – 30 degrees, to optimize the viewing angle (see Appendix 7B). This can be realized by pressing a button on the remote control. During the experiment a failure occurred in the television-picture or in the motorized swivel. This will be explained in more detail further on. With this experiment the UPFS of a failure in a low importance function and the UPFS of a failure in a high importance function were assessed. For both goals a different scenario was developed. See Appendix 7C for a detailed discussion.

Two homogeneous groups

For this experiment two groups were created: one group which tested the function with low importance and the other group which tested the function with high function importance. A choice had to be made for categorizing each participant. Two groups were needed. One group tested the high importance function. The other group tested the low importance function. The goal is to create two different groups which are equal except on the point of function importance. The group with the high importance function consisted of participants who rated watching the desired program, without failures in screen and sound as very important. For the other group with the low importance function, participants had to rate the use of the motorized swivel as not very important. Based on this method two groups were created significantly differentiating on the rating of function importance. In other words all participants within group one are a

homogeneous group and all participants in group 2 are also a homogenous group, but these two groups differ from each other on function importance score. Creating these two homogenous groups was done for controlling extraneous effects. Extraneous effects might influence the behavior being studied, but are not of interest to the researcher (Goodwin, 2005). This could lead to an erroneous interpretation of the results derived from the experiment. These two groups had their own scenario. The importance of the motorized swivel could change as the participant used the motorized swivel. Namely, participants could rate, after using it, the motorized swivel as (more) important (Gielen, 2007). To this important point will be returned later on.

Thinking aloud

The participant is also told to think aloud. With thinking aloud a large number of qualitative data can be collected from a fairly small number of users (Nielsen, 1993). The participant's comments often contain a lot of information which make the test results easier to interpret (Nielsen, 1993). Participants often make comments about things they like or do not like. Also informal comments about small irritants can be grasped with this thinking aloud protocol (Nielsen, 1993).

7.3 Pilot and redesign of experiment

For testing this experimental design, as described in section 7.2, five pilots were performed. After the first three pilots some remarkable things got clear.

- The scenarios had to be more directive.
- The function failure list had to be shorter.
- PANAS-X could not measure the caused irritation as expected

See appendix 7D for a detailed discussion.

For resolving these problems some adjustments were made. Two more pilots were done. The results with regard to PANAS-X were much more in line with the expectations. Negative affect did not decrease anymore. Another promising point is that the exit questionnaire showed that the person who had a picture failure perceived more irritation than the person who had a failure in the motorized swivel. Of course it should be noticed that in total only five pilots were done and only two of them with the new experimental design. So, the preliminary results were only a small indication of the direction the experiment would go.

Exit questionnaire

The exit questionnaire used in this experiment was first applied in an experiment conducted by the Quality and Reliability Engineering group (Technological University Eindhoven). In this experiment failures occurred in teletext. Participants had to answer on the propositions of the exit questionnaire. The reliability, Cronbach's Alpha, was sufficient: 0.78.

8. Results of the Experiment

In the previous chapter the (re)design of the experiment is outlined. The experiment was conducted as described in section 6.3. The results regarding the experiments are presented in this section.

In total 25 students participated in the experiment. Due to financial constraints only 25 people could be invited for the experiment[1]. Thirteen of these participants underwent the picture scenario and twelve participants underwent the motorized swivel scenario.

8.1 Control for irritableness with PANAS-X

The participants completed a PANAS-X before the "actual experiment" (one of the two scenarios) started. This PANAS-X measured the mood of the participants at the moment right now (PANAS-X-1). The mood at the moment right now of the participants could be very different for every particular student. However, students are a homogenous group, and in general the average mood in one group (scenario 1) should be equal with the mood in the other group (scenario 2). The means of general negative affect, basic negative emotion, Fear, Hostility, Guilt and Sadness between the participants of scenario one and two were compared. Care should be taken when interpreting the results with regard to Fear, Hostility, Guilt and Sadness, since the discriminant validity is questionable. The results are presented in table 8.1. A Mann-Whitney test was used for examining the difference in means (since the data was not normally distributed), figure 8.1.

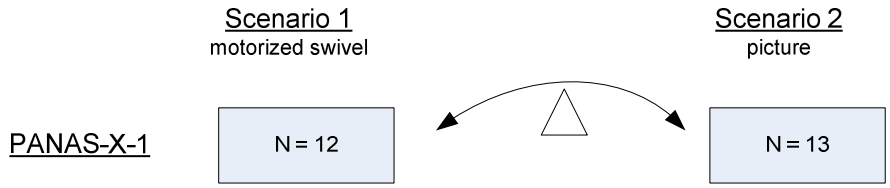


Figure 8.1: Examining differences between relevant PANAS-X scales

Although the participants describe their mood at the moment, which could be very different for every participant, there are no significant ($\alpha = 0.1$) differences found between the participants of scenario one and two.

Table 8.1: Mann-Whitney test PANAS-X-1 scenario 1 and 2

	Fear	Hostility	Guilt	Sadness	Gen Neg Affect	Bas Neg Emo
Mann-Whitney U	78.000	75.000	54.000	74.000	70.500	77.000
significance	1.000	0.894	0.205	0.852	0.689	0.979

So, based on the results above, PANAS-X can be used as a selection-tool for assessing with regard to irritableness level. The participants had in general all the same level of irritableness before the experiment started.

In section 6.3.3.2 is concluded that it is not possible to distinguish the basic negative emotions Fear, Hostility, Guilt and Sadness. These basic negative emotions stick to one emotion: basic negative emotion. In section 6.3. that the Basic Negative Emotion Scale is highly correlated with General Negative Affect

[1] All 25 participants of the experiment and 5 participants of the pilot get an electronic gadget worth € 25

(correlation: 0.916). Thus using the Basic Negative Emotion Scale in addition to General Negative Affect does not add that much extra information, both scales measure nearly the same constructs. Taking into account the easiness of use for participants, using the General Negative Affect scale alone would be recommended, since this scale yields the same information as using the Basic Negative Emotion Scale and the General Negative Affect scale. Returning to research question (I):

How to measure and control for "irritableness" of people interacting with a consumer product?

The answer is, use the General Negative Affect Scale of PANAS-X or PANAS. The General Negative Affect Scale of PANAS-X is the same as the General Negative Affect scale of PANAS.

8.2 Measuring UPFS with PANAS-X

UPFS is measured by means of PANAS-X-1 and PANAS-X-2. As argued in section 5.2 the relevant scales for measuring UPFS are: Fear, Hostility, Guilt, Sadness, General Negative Affect and Basic Negative Emotion. Care should be taken when interpreting the results with regard to Fear, Hostility, Guilt and Sadness, since the discriminant validity is questionable. For testing the significance of the differences of the means of the relevant scales of PANAS-X between the scenarios, several Mann-Whitney tests are done, see figure 8.2. For testing the significance of the difference within a scenario, a non parametric test for repeated measurers is done: Wilcoxon test. The result is presented in figure 8.2.

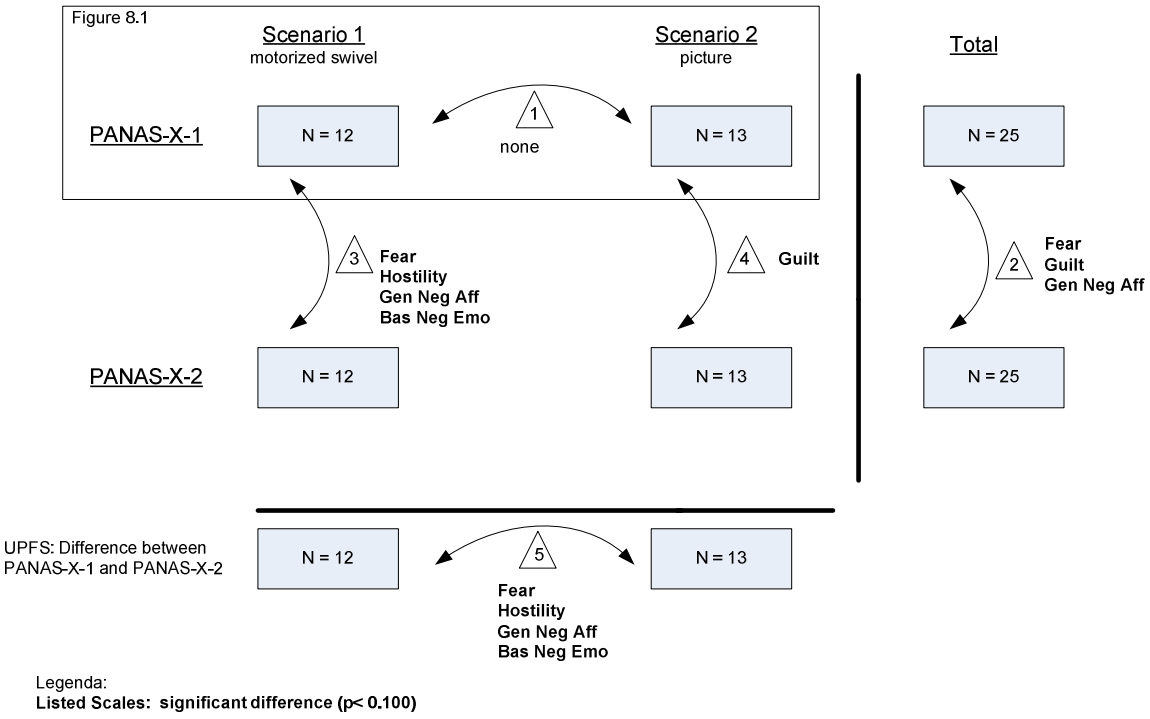


Figure 8.2: Testing for significance of the differences between PANAS-X and scenarios

The first Mann-Whitney test is described in section 8.1: testing the difference of PANAS-X-1 between scenario 1 and scenario 2. As described there were no significant ($\alpha = 0.1$) differences between scenario 1 and scenario 2.

For measuring UPFS there should be a difference between the scores of PANAS-X-1 and PANAS-X-2 (PANAS-X-2 is the PANAS-X the participants completed after the failure had occurred). The difference

between PANAS-X-1 and PANAS-X-2 is examined with a Wilcoxon test. No distinction is made between scenarios (N=25). In this second test are significant ($\alpha = 0.1$) differences between PANAS-X-1 and PANAS-X-2 in the relevant scales: Fear, Guilt and General Negative Affect.

The third and the fourth test (Wilcoxon tests) examine the difference between PANAS-X-1 and PANAS-X-2 for both scenarios individually. In scenario 1 the differences of the scores on the scales Fear, Hostility, General Negative Affect and Basic Negative Emotion are significant. Only Guilt can measure a significant difference in score before and after the experiment in scenario 2. As said before UPFS is defined here as the difference between PANAS-X-1 and PANAS-X-2 on the scales General Negative Affect and Basic Negative Emotion. These differences are significant in scenario 1 but not in scenario 2. So, it is not possible to conclude that the occurred picture failure (scenario 2) caused (more) irritation (a higher level of UPFS). However it is interesting to check whether there is a difference between the significant UPFS in scenario 1 and non significant UPFS in scenario 2.

Therefore a Mann-Whitney test (test five) is performed for testing if the UPFS between the two scenarios differs. There is a significant difference in the difference of Fear, Hostility, General Negative Affect and Basic Negative Emotion between the scenarios. Based on these results there is a strong indication that the UPFS in scenario 1 (the motorized scenario) is larger than the UPFS in scenario 2 (the picture scenario), see also appendix 8A.

With the results presented above a preliminary answer to research question (II) can be given.

How to measure User Perceived Failure Severity?

PANAS-X could measure UPFS in the motorized swivel scenario but not in the picture scenario. So, PANAS-X should not be used for measuring UPFS.

8.3 Measuring UPFS with the exit-questionnaire

In this research project, the reliability of the exit questionnaire is not yet assessed. The underlying structure of the exit-questionnaire is examined in this section with a correlation matrix, table 8.2. This underlying structure is examined because the exit questionnaire consists of two constructs: emotional response and action response. The items describing emotional responses are related to irritation and so related to UPFS. In Appendix 8B the structure of the exit questionnaire is given.

Table 8.2: Correlation matrix items exit questionnaire

	enormous irritating	go to service-desk	become enraged	just wait	not angry	return tv back	searching at Internet	small failure	ask relative for help	does not bother me
enormous irritating			**.61****	.48*	**.51****			**.57****		**.66***
go to service-desk								.41**		
become enraged	**.61****			.50*	**.40***	.64**		**.53****		**.69***
just wait	.48*							.64**		
not angry	**.51***		**.40***							
return tv back			.64**							.51**
searching at Internet										
small failure	**.57****	.41*	**.53****	.64**						**.42***
ask relative for help										
does not bother me	**.66****		**.69****			.51**		**.42***		

** Correlation significant at the 0.01 level

* Correlation significant at the 0.05 level

This correlation matrix (table 8.2) shows that the items "searching at Internet" and "asking a relative for help" are not related with any other item. Also the item "go to service-desk" is only related with one other item: "(this failure is a) small failure". These items are all action responses. Apparently don't have these items much in common with the other items, also not with the other action response items. This could be an indication that the construct action response is not measured well with this exit questionnaire. The items related to emotional response correlated very high (bold numbers table 8.3). The reliability of the exit questionnaire, Cronbach's alpha, is shown below, table 8.3. The reliability increases considerably when these three items are deleted. The exit questionnaire with 7 items had a reliability of 0.842 which is quite good.

Table 8.3: Cronbach's alpha exit questionnaire

Cronbach's alpha		
10 items	0.77	all items
9 items	0.79	all items except: ask relative for help
8 items	0.82	all items except: ask relative for help, searching at Internet
7 items	0.84	all items except: ask relative for help, searching at Internet, go to service desk

The reliability of the two individual the scales: emotional response and action response is also assessed. The reliability of the emotional scales is quite good 0.83; this score is well above the threshold value of 0.7. As expected, the reliability of the action response scale, 0.34, is far below the recommended value of 0.7. The score's related to the emotional response are at interest when measuring UPFS. Therefore the sum of the individual scores on the five items related to emotional response (see also Appendix 8B) are presented in table 8.4.

Table 8.4: Individual scores emotional response scale

Scenario 1 motorized swivel														
Emotional Response score (UPFS) of participants (N=12)													average score	st dev
22	17	22	18	15	14	23	18	21	18	22	24		19.50	3.26
Scenario 2 picture failure														
Emotional Response score (UPFS) of participants (N=13)													average score	st dev
18	25	22	22	9	25	23	24	23	15	20	25	23	21.08	4.66

In scenario 2 there is one participant with a very low score in comparison with the other participants of this scenario. This outlier is also visible in the box plot below, Appendix 8C.

Outliers can be classified in four categories based on the source of uniqueness, Appendix 8D. The outlier in this experiment can be classified as an extraordinary observation. The reason is that there are no particular circumstances which are different in comparison with other experiments. Nevertheless this results in a very low emotional response score. The average score and standard deviation is heavily influenced by this participant. Removing this outlier decreased the standard deviation in scenario 2 from 4.66 to 3.06. The average irritation score in scenario 2 increased to 22.08 from 21.08, table 8.4 and 8.5. Based on this heavy influence on the average score and standard deviation this outlier was removed.

With the reaming of the dataset (N=24) a non parametric test was done for comparison of means. The result is that the participants in scenario 2 (the picture scenario) had a higher irritation score than the participants in scenario 1. The difference between scenario 1 and scenario 2 was significant ($p = 0.04$).

Based on this it could be concluded that scenario 2 (the picture scenario) caused more irritation than scenario 1 (the motorized swivel scenario).

Table 8.5: Individual scores emotional response scale (without outlier)

Scenario 1 motorized swivel													
Emotional Response score (UPFS) of participants (N=12)												average score	st dev
22	17	22	18	15	14	23	18	21	18	22	24	19.50	3.26
Scenario 2 picture failure													
Emotional Response score (UPFS) of participants (N=12)												average score	st dev
18	25	22	22	25	23	24	23	15	20	25	23	22.08	3.06

Based on the results above, it is shown that the emotional response scale is reliable and can be used for measuring UPFS. This will answer research question (II)

How to measure User Perceived Failure Severity?

User Perceived Failure Severity can be measured with the emotional response scale.

Based on this scale the conclusion can be drawn that a failure in a high importance function leads to more perceived failure severity than a failure in a low importance function.

8.4 Other results

Besides the results given above, there are also some other results worth mentioning. These will be presented below.

8.4.1 Failure Impact Picture scenario

As given in figure 2.1, failure impact could influence UPFS. The influence of failure impact on UPFS was not examined within this experiment. Therefore controlling for failure impact was necessary within this experiment. Whether this succeeded can be checked with the function failure list. The (average) score should be equal for both scenarios. In table 8.6 the severity scored of the failure of both scenarios is given. Scenario 1 had a failure impact severity score of 3.67 in comparison with a score of 2.36 for scenario 2. The failure impact severity of both scenarios is significant different ($p = 0.04$). The effect on UPFS could be that in this experiment the measured UPFS in scenario 2 is too low in comparison with scenario 1. It could be expected that when the failure impact severity of scenario 2 is higher (as high as scenario 1) also the UPFS would increase. An explanation for the fact that the failure impact in scenario 2 (picture scenario) was not that high as in scenario 1 (motorized swivel scenario) is given below.

With regard to the picture scenario, the picture quality of the television in the normal situation during the experiment was not perceived by the participants as good. This was caused by the device developed for creating the picture disruption. In the normal situation (good picture quality) this device disturbed the signal a little. So, the picture was a little bit less good than normally could be expected. However another problem is the size of the television. The television used in this experiment had a very large picture diagonal (107 cm). So, the number of pixels per cm^2 is low compared with other smaller televisions. This will reduce the picture quality as well. Also in combination with the small viewing distance, the picture quality seems a little bit less good. This is also reflected in the function evaluation list. In the motorized swivel scenario participants did not only mention the failure in the motorized swivel but also saw the perceived picture quality as a failure (table 8.6). In table 8.6 are all failures listed that are mentioned by the participants.

Table 8.6: Failure impact severity.

scenario 1	function	times mentioned	average score**
correct failed function	**motorized swivel**	**12**	**3.67**
other functions	watching desired program	6	2.33
	watching two programs	3	2
	teletext	2	1
	changing picture format	1	3
scenario 2	function	times mentioned	average score
correct failed function	**watching desired program**	**11***	**2.36**
other functions	motorized swivel	4	1.25
	teletext	2	2.5
	changing picture format	2	1.5

*One participant is deleted due to the abnormal score on the exit questionnaire (see table 8.7). Another participant did not perceive any failure in the picture.
** A scale of 0 to 5 was used.

From table 8.6 it becomes apparent that participants from scenario 1(motorized swivel) perceive the failure impact in the picture with a severity of 2.33. Despite that there should be no failure in the picture in scenario 1. However this perceived picture failure in scenario 1 is nearly as high in scenario 2 (the picture scenario). In the "real" picture failure scenario (scenario 2) participants perceive the failure with a severity of 2.36 (table 8.6). So, the failure in the picture scenario did (nearly) not worsen the failure severity. This could be a reason why there is no significant difference between PANAS-X-1 and PANAS-X-2 in scenario 2, see figure 8.2.

8.4.2 Causal attribution

Beside the results presented above also some other results are worth mentioning. In the picture scenario, participants did not blame the television for bad picture quality. In the table below (8.7) several reactions are listed on the question what caused the bad picture? The participants did not tend to blame the television itself. Apparently the participants do not think that a picture failure could occur in a television as used in this experiment. Participants attributed the cause of this picture failure in most cases (10 of 13) to things that are not the television itself, for example the cable company or the weather.

Table 8.7: Some reactions on the question: What caused the bad picture? (N=25)

"I don't think that the television is the problem, probably the signal is bad"
"Probably it isn't the television who causes the bad picture"
"There was somewhere a short disruption, the problem has to do with the weather I guess "
"The signal is bad"
"The fine-tuning is not good enough"
"The antenna is not properly connected"
"Maybe the cable is the problem"

In this situation participants evaluated the severity of the picture failure with an average score of 2.36 (table 8.5), despite the fact that this function is a high importance function. In comparison the failure in the motorized swivel, a low importance function, was evaluated more severe with an average score of 3.67 (table 8.5). Since the expected relation ship of failure impact UPFS, this finding could be an explanation why the UPFS in scenario 1, the low importance function, is larger than the UPFS in scenario 2 the high importance function, see figure 8.2 and Appendix 8A.

Circumstances related to the frustrating incident and the situation of the person itself can lead to a difference in the level of frustration experienced by that person (perceived failure severity) (Lazar, 2006). Persons will have a high commitment to a goal when the goal is important to them and they believe that

the goal can be reached (Locke,1996). The level of frustration that people experience would be influenced by how important the goal was to them, as well as how confident they are in their abilities (self-efficacy) (Lazar, 2006). Self-efficacy can be seen as how well a task can be performed when it involves setbacks, obstacles, or failures (Bandura, 1986). In the picture scenario (scenario 2) the task, the goal: "describe the broadcasted program", could be (partly) completed despite the fact that the picture quality was heavily decreased. It was still possible to recognize several programs. Whereas in the motorized swivel scenario the task, the goal: "turn the television", could totally not be completed.

There are also other forces that may influence the force of the frustration. Not all obstructions are equally frustrating (Lazar, 2006). Persons can perceive the "failure" as justified by socially accepted rules, in this case the frustration response may be minimized (Baron, 1977). Extra information available to the individual may reduce the expectations. At times every person has experienced a cable disruption at home. It is known that this could happen. The failure in the picture in scenario 2 could be associated with this socially accepted phenomenon of cable disruption. According to Baron this will lead to a lower level of frustration (perceived failure severity) (Baron, 1977). In the case of the motorized swivel it is not accepted that the motorized swivel will fail (once in a while). The perceived failure severity is higher in comparison with the perceived failure severity of the picture failure. This is also supported in the paper of Westbrook and Oliver (1991). They state that when an evaluation of the relevant consumption experience (or its associated product or service) is required, past experiences and memories such as prior expectancies, and disconfirmation beliefs play a role in this evaluation (Westbrook, 1991). In this experiment these prior expectancies could be twofold. The first one is the overall expectancy of a television. Past usage and buying experience of a television created the expectancy that a television always should have at least "good" picture quality. So, the participants could not "believe" that, based on their past experience, the bad picture quality was caused by the television. The second point is the image of Philips. The television used was a Philips LCD television type 42PF9830/10. The (brand) image of Philips is very good (Interbrand, 2006). For example "The high quality, modern designs, and overall good image of this well-known brand [Philips] impressed Chinese consumers" (Lee, 2001). Participants do not expect that a product such as this €3000 worth LCD-TV could fail on such an important function. Individuals are cognitive misers (Taylor, 1980): individuals are liable to accept the easiest explanation (Hansen, 1980). The easiest explanation is a failure in the cable, and not a failure of the television.

These results can be visualized in a matrix, figure 8.3, the UPFS-matrix. The user perceived failure severity depends on the function importance and the failure attribution. The first quadrant of UPFS matrix is in the situation where a high importance function fails and the failure is attributed externally. The existence of this quadrant is proven with the function failure list. Also the results from PANAS-X with regard to UPFS indicate that the severity of a low importance function internally attributed is perceived higher by participants than a failure in a high importance function externally attributed (see figure 8.2 and Appendix 8A). From table 8.4 however, when explicitly was told that the failure was caused by the television (internally) (quadrant 2) participants rated the severity of this failure higher (22.08) than a low importance function internally attributed (quadrant 3) (19.50). Quadrant 4 is not examined in this master thesis project. Based on the results it would be expected that an externally attributed low importance function would be perceived less severe than a high importance externally attributes function. However the existence of this category failures is questionable. It would be difficult to attribute failures, like a not working motorized swivel or ambilight, to other things than the television itself.

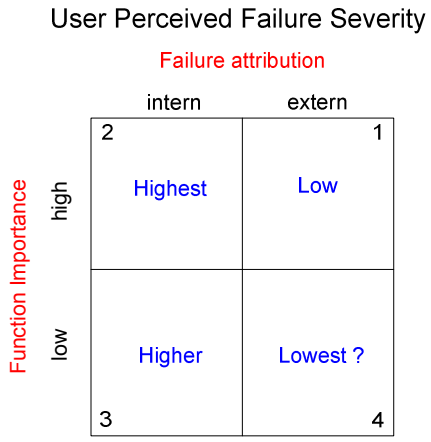


Figure 8.3: User perceived failure severity matrix, UPFS-matrix

One major limitation is that the results presented in sections 8.1, 8.2 and 8.3 are based on a small sample size (N=24). Generalizing these results should be done with care. The results obtained are an indication of the severity users perceive when a failure occurs.

8.4.3 Further research direction

One limitation of the previous experiment is the limited sample size (N=25). For validating the UPFS-matrix, a 2 by 2 factorial experimental ANOVA design is needed, with a larger sample size. The 2 by 2 factorial design is constructed as follows: function failure at a low and high a level, and failure attribution at two levels: intern and extern. The recommend sample size depends on: alpha (α), the power (β), effect size, and number of groups. The general recommended level for alpha (α) is 0.05 and for the power (1-β) 0.80. The number of groups is 4, see figure 8.3. There are no other data available for calculating the effect size except the data presented in table 8.6. These data are only related to 2 groups: one group, low function importance and intern failure attribution and the other group: high function importance and extern failure attribution. These data will be used for calculating the effect size as if the experiment could be analyzed with one-way ANOVA: one independent variable, function importance and one dependent variable, UPFS. The effect size is calculated at 0.38 see appendix 8E.

The effect size of 0.38 should be interpreted with care because the method used to calculate this value is based on a one-way ANOVA. One of the assumptions underlying one-way ANOVA is the assumption that the distribution of the dependent variable is normally distributed. This is not the case. However with a moderate or larger sample size, even when the normality assumption is violated, one-way ANOVA will yield reasonable results (Green and Salkind 2002). But a sample size of 12 per group can not be considered as moderate (or large). The result could be that the power may be reduced considerably if the population distributions are non-normal, i.e. thick-tailed or heavily skewed (Green and Salkind 2002). The data presented in table 8.5 is not particular thick-tailed or heavily skewed, but is certainly not normal distributed. Due to the lack of other data the best estimate is an expected effect size of 0.38. According to Cohen this effect of 0.38 can be considered as large (0.40 = large, 0.25 = medium, 0.10 = small). Using the statistical program Power and Precision (Borenstein, 2001) for calculating the recommend sample size of a 2 by 2 ANOVA, with the values of α=0.05, 1-β = 0.80 and a large effect size = 0.40, results in a recommended sample size of N = 56 (28 cases per level). As stated above when the normality assumption is not met, the power of the ANOVA test may be reduced heavily. Since this threat is real in this situation a solution would be to design the experiment with a higher power level so that a decrease in power by non-normality still would yield an acceptable power level. In appendix 8F, the power as a function of sample size (per group) and effect is given. The result is that for a higher power level a larger sample size

is needed. The side effect is that that the effects of non normality will be reduced extra since a larger sample size will yield more accurate results (increase in power) even when the normality assumption is violated (Green and Salkind 2002).

8.4.4 Conclusion

From section 8.1 it becomes clear that the Negative Affect scale of PANAS-X can control for irritableness but that PANAS-X can measure UPFS can not be fully proven. It has to be noted that this conclusion is influenced through failure impact. The impact of the failure caused by the bad picture quality (the failure situation) is not almost worse than the normal situation (no failure). Therefore it could be expected that the difference (UPFS) between PANAS-X-1 and PANAS-X-2 is not significant. In an experiment with better possibilities for controlling failure impact, PANAS-X could measure UPFS. In section 8.2 is showed that the emotional response scale can measure UPFS. UPFS depends on function importance and causal attribution. A failure in a high importance function internally attributed (the interaction of user with television) will be perceived more severe than a failure in a high importance function externally attributed (beyond the interaction of user with television). Another result is that a failure in a low importance function internally attributed will be perceived more severe than a failure in a high importance function externally attributed. This relation is given in the UPFS-matrix, figure 8.3. The results presented in the UPFS matrix need to be interpreted with care due to the low sample size. An experiment that would investigate the proposed relations in the UPFS-matrix should use a larger sample size. Also a better control for failure impact is necessary.

9. UPFS Matrix

Based on the results of the previous chapter a revised research model will be presented (section 9.1) and a new research question will be introduced. Based on the results of chapter 8 several hypotheses are given with regard to the research question. The revised research model and the accompanying research questions and hypotheses will be tested by means of an experiment.

9.1 Revised research model

Based on the results of chapter 8 the research model presented in figure 1.4 and figure 1.5 needs to be revised. Failure attribution plays an important role in the user perceived failure severity. Attributing the failure internally or externally influenced the UPFS heavily. An internal attributed failure leads to a higher level of UPFS than a failure external attributed ceteris paribus. Intern is defined here as the interaction of the user with the television. Extern is everything beyond the interaction of the user with the television, see also appendix 9A. Also the importance of the failure influences the UPFS. It is difficult to direct the influence of failure impact. Therefore failure impact is removed as independent variable. Failure impact is will be treated as an extraneous variable. The goal is to keep the level of failure impact as constant as possible during the experiment. The results from chapter 8 reveal that another procedure is needed as used in the previous experiment, because the failure impact was not equal across the scenarios, see table 8.6. Also the irritableness of people needs to be kept constant during the experiment. As said before a disproportionate distribution of irritableness levels may reduce the relationship between the dependent and independent variable. See for the revised research model appendix 9B.

9.2 Research questions

The relationships given in the research model presented in the previous section can be translated into the following research question (III):

III. ***To what extent do Failure attribution (FA) and Function Importance (FI) influence UPFS?***

The results presented in chapter 8 and the accompanying UPFS-matrix are the starting point for testing the research model as given in figure 8.3 and answering research question (III). The expected relationships between failure attribution, function importance and UPFS are based on the results in chapter 8 and are hypothesized as follows:

H 1A: The highest UPFS will occur in a high importance function internally attributed.

H 1B: The lowest UPFS will occur in a low importance function externally attributed.

H 1C: The UPFS of a low importance function internally attributed is higher than the UPFS of a high importance function externally attributed, respecting *H 1A* and *H 1B* .

These hypotheses will be confirmed or rejected by means of an experiment. Consequently an answer on research question III can be given. As given in section 8.4.3 a 2 by 2 factorial design needs to be developed. One important limitation of the previous experiment is the limited sample size. As calculated in 8.4.3 a sample size larger than 56 is needed. For validating the UPFS-matrix, four different scenarios are needed, one for each quadrant in the UPFS-matrix, see figure 8.3. For this experiment also a convenience sample of students will be used since students are considered to be a sound, homogenous

respondent group in consumer research (Peterson, 2001). To the experimental design will be returned later on.

9.3 Failure Attribution

In this section more information about failure attribution will be given.

Failure attribution (FA) is related to the concept of causal explanation. Heider (1995) was one of the first to explain individuals' reactions to other individuals or environment such as weather or economic conditions. As already said, individuals are cognitive misers (Oliver 1996): individuals are prone to accept the easiest explanation for events. In their judgment about action individuals are biased in the attribution of their actions. Attributions for the same outcome are differentially assigned depending on whether the outcome is attributed to self or to others. This phenomenon is called the fundamental attribution error: "There is pervasive tendency for actors [the self] to attribute their actions to situational requirements, whereas observers [of others] tend to attribute the same actions to stable personal dispositions" (Jones and Nisbett, 1972). Attributions to environmental events are called situational attributions. Attributions to stable characteristics of people are called dispositional attributions. Positive outcomes of actions are perceived with the perception of egocentric bias and lead to self-serving attributions (Oliver 1996). This means that individuals write success to their own characteristics. Other's successes are attributed to situations. This tendency reverses when explanations for negative events are needed. For protecting their ego individuals ascribe their failures to situational reasons and others' failures to the characteristics of people (Oliver 1996).

Weiner has elaborated on Heider's original work. Weiner's attribution framework has three dimensions: Internal/External, Stability and Controllability (Weiner 1985a). The internal/external dimension is referred to as the locus of causality and is the same as the dispositional/ situational difference. This dimension is relevant in this research because people tend not to attribute the failure internal (interaction between user and television) but tend to attribute the failure external in the high importance function scenario as opposed to the low importance function scenario. The second dimension is stability. Some internal causes can be considered stable and predictable (like skills) while others are variable (study effort) (Oliver 1996). This dimension signals whether the same failure can be expected in the future or whether the failure one-time event (Oliver 1996). The third dimension is controllability. Controllability examines if variables can be modified by the actor or an external agent. For example, effort is controllable and aptitude not (Oliver 1996). These two dimensions are not at interest in this master thesis, but could influence the relation between FA and UPFS. Therefore controlling for stability and controllability is necessary.

Anything unusual which stimulates individuals' attention to the outcome will lead to causal search (Oliver 1996). Disconfirmation of expectations will cause attribution processing. This disconfirmation of expectations can be categorized into: unexpectedness, unattained or frustrated goals (including negative outcomes), or other like unusual success or outcomes of great importance (Weiner 1985b). Particular the second category, unattained of frustrated goals, plays in an important role in the performed experiment (chapter 7). Participants could not complete their goals: turn the television with the motorized swivel or could not list the programs due to the bad picture quality.

In this chapter the adapted research model and the accompanying research question with hypotheses were presented. Also some theoretical issues about failure attribution were given. The next chapter will continue with the design of the new experiment.

10. Second experiment

The experimental design will be presented in section 10.1. The goal of this experiment is to test the hypotheses presented in section 9.2. The experimental design will be explained in this section. In section 10.2 the results of the pilot will be presented.

10.1 Experimental design

First a comment will be made about the difference between this (second) experiment and the previous experiment.

Comparison with previous experiment

This second experiment was largely the same as the previous experiment described in chapter 7. However one important difference was that in the previous experiment participants perceived a failure situation of a television. A living room setting was created and the home conditions were copied as good as possible. Despite, it was not possible to create exactly the same conditions as home. Participants knew that they participate in an experiment and it could be expected that they behave differently than in their "normal" everyday surroundings. The previous experiment was very time consuming and expensive. Therefore only a small number of people could participate (N = 25). The advantage was that a lot of qualitative information was gathered about consumers in failure situations. In this second experiment participants did not "perceive" the failure as in the previous experiment. As already said it was not possible to create the "normal" everyday surroundings. Therefore the time consuming design of the previous experiment was relaxed. An experimental design was chosen where it was possible to perform the experiment with a larger sample size. This enlarged the validity of the results. In the previous experiment participants had to do a specific task, listing programs or turning the television with the motorized swivel, and during this task they were confronted with a failure. In contrast, in the second experiment the failures (four different scenarios on for each failure category of the UPFS-matrix) were shown to the participants (by means of film and slides) and then these participants had to "evaluate" these failures on user severity. This experimental design was very mobile. Therefore the experiment could be performed at places were a lot of possible participants were available. For this experiment also a homogenous group of students was needed for controlling extraneous effects. For the design of the experiment see Appendix 10A.

10.2 Results of pilot

Before the actual experiment started some pilot sessions were held. The results of these sessions led to some adaptations of the experiment. These results and adaptations will be discussed below:

Too much information

The movie about the introduction of the television and its functions contained too much information. Therefore participants could not grasp all information. Participants had to pay attention at the film itself. Besides that beneath the film there was some textual information what they also had to read. This text contained information about what kind of function was showed at the film, what you could do with the function and how you could invoke these functions by using the remote control. On the right hand of the film the remote control was shown and for each function the accompanying button was highlighted. It was not clear to the participants where they had to focus their attention on. The most important information the participants should grasp was the film itself and the accompanying text about what kind of function was showed. Therefore, the text about how you could invoke the functions and the picture of the remote control were eliminated.

The start of the movies
The announcement of the films occurred at the end of the "PANAS"-form and the "Function Importance Questionnaire"-form. The placements of these announcements were not that good. After the participants completed these two (questionnaire)-forms they lost attention and pushed the "next" button. So, the start of these films was a surprise for them. At the time the participants noticed that they had to pay attention to the film they already missed some valuable information. Therefore a separate form was made for the announcements of the films.

Programming errors
During the pilot it became apparent that not the right questions were showed to the participant. This failure was caused by a programming error and was solved when the actual experiment started.

Failure attribution questionnaire
The failure attribution questions could be clearer according to the participants. Some questions (question 2 and 3, appendix 10B) could be formulated simpler. Therefore a revision of this questionnaire was made. However this revision should stay as close as possible to the original version of this questionnaire. Because, the original Failure Attribution Questions are based on the Russell's Causal Dimension Scale (Russel 1982). This scale is internally consistent and reliable ($\alpha = 0.867$). A too large deviation from these original questions does not guarantee this internal consistency and reliability. The new Failure Attribution Questions are presented in appendix 10C.

11. Results of the Experiment

In the previous chapter the (re)design of the experiment is outlined. The experiment was conducted as described in section 10.1. The results regarding this experiment are presented in this section.

In total 149 students participated in the experiment. Thirty-five participants underwent scenario 1, 37 scenario 2, 38 scenario 3 and 39 scenario 4, see also appendix 11A.

Before the results will be presented some remarks have to be made with regard to the experiment. After 13 participants underwent the experiment some "unexpected" results appeared with regard to the failure impact questionnaire. According to the participants it was not clear to them that this question pointed explicitly at the situation were the function failed. Therefore the question was changed. However the failure impact questionnaire was still not clear enough (for the participants 14 through 44). The problem was that the worth "impact" was ambiguous. Participants interpreted this worth differently. Therefore the worth impact was replaced. See appendix 11B for this revised questionnaire. Besides this, a programming error caused the situation that 4 participants get the wrong Failure Impact Question. Therefore these 4 values were deleted in the database.

A lesson learnt is that doing consumer research as done in this second experiment (automatic procedure) requires an extensive test period with real participants. This is necessary for preventing technical program errors and difficulties with questionnaires. During this automatic procedure it is not possible to make adjustment for solving possible problems ad hoc. During the first experiment in a laboratory setting there is more flexibility to solve possible problems. Therefore the test period of an automatic procedure experiment needs to longer than for an experiment in a laboratory setting.

11.1 Examination of the experiment

In this chapter will be examined whether the participants perceived the scenarios as intended (described in section 10.1) by the researcher. Results with regard to PANAS, Failure Impact Questionnaire, Failure Attribution Questionnaire and Function Importance Questionnaire will be discussed. For this discussion some results obtained by ANOVA will be used. See Appendix 11C for the assumptions underlying ANOVA.

11.1.2 Irritableness

As said before a disproportionate distribution of irritableness levels may reduce the relationship between the dependent and independent variables. There will be controlled for irritableness by use of the Negative Affect dimension of PANAS. Therefore the reliability (Cronbach's Alpha) of the PANAS-scales are calculated, see table 11.1.

Table 11.1: Reliability Analysis

Scale	Cronbach's Alpha
Negative Affect	0.80
Positive Affect	0.81

The reliability of both scales is good ($\alpha \geq 0.80$). Only the Negative Affect scale is at interest for controlling for irritableness.

The data of the previous experiment with regard to PANAS (N=200) and this second experiment (N=149) are combined. Based on this sample size (N=348)[2] the average Negative Affect score is calculated: 18.69 and the accompanying standard deviation: 5.26. So, the acceptable irritableness level is 13.43-23.95 (18.69 +/- 5.26). Based on these levels 36 people of the 149 had to be deleted from the database. Fifteen participants scored below the lower threshold value and 21 participants scored above the upper threshold value. This resulted in a sample size of N=113, distributed as given in figure 11.1.

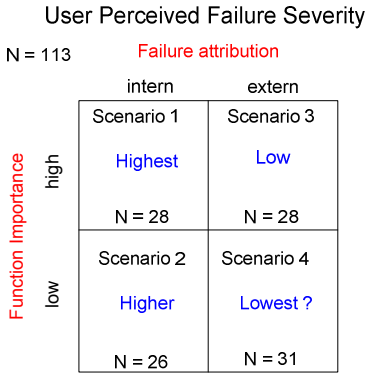


Figure 11.1: Number of participants in each scenario after controlling for irritableness

These 113 participants - having an average level of irritableness (group 2)- will be compared with the other two groups: the participants with a low irritableness level (group 1)and the participants with a high irritableness level (group 3). With this comparison will be checked if the level of irritableness (measured by Negative Affect) really influences the dependent variable UPFS (measured by ERS-score). For this comparison one-way ANOVA could be used.

Since the assumptions regarding ANOVA are not met, performing a Kruskal-Wallis test is recommended (Hair et al. 2005), see Appendix 11D for a detailed discussion.

Also here the reliability of the Emotional Response Scale has to be checked. Cronbach's Alpha = 0.80. Therefore the reliability of the Emotional Response scale is good ($\alpha \geq 0.80$).

Comparing these 113 participants - having an average level of irritableness- with the other two groups: the participants with a low irritableness level and the participants with a higher irritableness level, led to some remarkable results. In table 11.2 the different scores on the Emotional Response Scale (ERS), a measure for UPFS, are given. From the Kruskal-Wallis test could be concluded that the difference in population medians of the ERS-score are marginally significant p=0.05. More specifically a follow-up test showed that the difference in ERS-score between group 1 and 2 is significant (p=0.02). As expected the ERS-score of group 3, (table 11.2), is higher than the ERS-score of group 2, but not significant. However, an unexpected result is that the ERS-score of group 1 is higher than ERS-score of group 2 and even higher than group 3, see table 11.2. Also the ERS-score of group 1 has a very low standard deviation in comparison with the other groups.

Since, there is a difference in ERS-score (a measure for UPFS) between groups, there can be concluded that it is necessary to control for irritableness when measuring UPFS.

[2] One participant filled in PANAS in both experiments and is therefore deleted in one of the databases.

Table 11.2: Average ERS-score of the different groups

Group*	N	avg ERS-score	SD ERS-score
1	15	20.73	2.84
2	113	18.12	4.17
3	21	19.33	4.12

Group 1= participants with a PANAS Negative Affect score below 13.43
Group 2= participants with a PANAS Negative Affect score between 13.43 and 23.95
Group 3= participants with a PANAS Negative Affect score higher than 23.95

The remaining of the analysis will continue with only group 2 (N=113) as given in figure 11.1.

11.1.3 Failure Impact

By the results of the Failure Impact Questionnaire it is possible to examine whether the participants perceived the same failure impact across the different scenarios. The experiment was designed to perceive a 100% loss of functionality of the failed function. As noted earlier there were during the experiment some problems with regard to the Failure Impact Questionnaire. In the case the participants did not perceived the same failure impact in the different scenarios, a wrongful conclusion could be drawn with regard to the effect of the treatment on UPFS. For this comparison one-way ANOVA could be used. The sample size per cell is moderate (ranging from 25-31) but the violations of the normality assumptions are that hard that also a non-parametric test, a Kruskal-Wallis Test will be done. Appendix 11E for a detailed discussion.

The average Failure Impact Questionnaire score is given in table 11.3 below. Based on the Welch statistic and Brown-Forsythe statistic, the hypothesis that the FIQ-scores are the same between the different scenarios cannot be rejected (p=0.23 and p=0.34 respectively). Also the non parametric Krusal Wallis Test did not find a significant difference between the FIQ of the different scenarios (p=0.30). Based on these results can be concluded that there is reasonably well controlled for Failure Impact.

Table 11.3: Average Failure Impact Questionnaire Score

Scenario*	N	avg FIQ-score
1	25	3.44
2	27	3.41
3	26	2.85
4	31	3.32

Scenario 1: Failure Importance high, Failure Attribution intern
Scenario 2: Failure Importance high, Failure Attribution extern
Scenario 3: Failure Importance low, Failure Attribution intern
Scenario 4: Failure Importance low, Failure Attribution extern

11.1.4 Failure Attribution

A Failure Attribution Questionnaire was added for controlling if the participants perceived the cause of the failure internally or externally. In the most ideal situation the score in the Failure Attribution Questionnaire (FAQ) should be 3 in the intern situation and 15 in the extern situation. This is not the case, see table 11.5. At least there should be a significant difference in FAQ score between the internal and external situation. One-way ANOVA could be used for comparison between groups. Based on a assumption check it is justified to use ANOVA, see appendix 11F for a detailed discussion.

There is a significant difference in FAQ-score ($p < 0.001$, for the Welch test and Brown-Forsythe test) between the internal and the external situation. As intended, the internal situation scores more “internally”

(a low score on the FAQ) than the external situation where a high score on the FAQ is expected, see table 11.4.

Table 11.4: Average Failure Attribution Questionnaire Score

Group*	N	avg FAQ-score
1	54	6.39
2	59	11.08

Group 1: Failure attribution intern (scenario 1 and 3)
Group 2: Failure attribution extern (scenario 2 and 4)

A more detailed examination of the differences in FAQ between the different scenarios is given below.

Table 11.5: Differences between FAQ-score

Scenario*	N	Avg FAQ-score	scenario	1	2	3	4
1	28	6.71	1		x		
2	28	13.79	2	x		x	x
3	26	6.04	3		x		x
4	31	8.65	4		x	x	

x = significant differences ($p<0.05$) in FAQ-score between scenarios
Scenario 1: Failure Importance high, Failure Attribution **intern**
Scenario 2: Failure Importance high, Failure Attribution **extern**
Scenario 3: Failure Importance low, Failure Attribution **intern**
Scenario 4: Failure Importance low, Failure Attribution **extern**

In table 11.5 the average FAQ-score is given for the different scenarios. The cells with a "x" mean that there is a significant difference between the scenarios. In the green cells a significant difference is expected. The intern scenarios should differ significant from the extern scenarios. The intern scenarios (1 and 3) should not differ significantly from each other, because the cause of the occurred failure was in both scenarios the same: a software error. Also the extern scenarios (2 and 4) should not differ significantly from each other, because in both scenarios the cause of the occurred failure was external: an error at the cable company (scenario 2) and an error at the power station (scenario 4).

The FAQ-score of scenario 1 does not differ significantly with the FAQ-score of scenario 3. The FAQ-score of scenario 1 differs significantly with the FAQ score of scenario 2 but not with scenario 4. The FAQ score of scenario 3 differs significantly with the FAQ-scores of scenario 2 and 4. The FAQ-score of scenario 2 differs significantly with the FAQ-score of scenario 1 and 3, but also with scenario 4. The FAQ-score of scenario 4 differs significantly with the FAQ-score of scenario 3, but not with the FAQ-score of scenario 1. Also, differs the FAQ-score of scenario 4 significantly with scenario 2.

Conclusively, the design of the intern scenarios (1 and 3) with regard to failure attribution is sufficient. The design of the extern scenario 2 is also sufficient, it differs significantly of the intern scenario 1 and 3. However the design of extern scenario 4 could be better. There is not enough contrast with intern scenario 1 and differs too much of extern scenario 2. The perceived failure attribution of scenario 4 is not "external enough". In scenario 4 the motorized swivel failed but the (high important) function "watching the desired program" was undamaged. Participants could think that it was not likely that during a power disruption one function ("watching the desired program") could work and the other function ("motorized swivel") could not work. Therefore the participants could think that despite the power disruption the failure of the motorized swivel is caused by the television itself.

11.1.5 Function Importance

In table 11.6 can be seen that for high important functions like watching the desired program the standard deviation is much smaller than for low important functions like Ambilight, Dual screen or the motorized swivel. Apparently some users do not see any added value of some functions while others users see on the other hand added value in some functions.

Table 11.6: Function Importance scores

	Watching desired program	Ambilight	Dual screen	Motorized swivel
Mean	9.50	4.81	5.30	4.51
SD	0.69	2.04	2.20	2.28

Based on the results above it can be concluded that there is reasonably well controlled for Failure Attribution and Failure Impact. The design of scenario 4 could be improved. Also it has been showed that it is necessary to control for "irritableness" when measuring UPFS.

11.2 Hypothesis testing

In this section the hypothesis as presented in chapter 9 will be tested. At the end an answer will be given at the research question. The hypotheses are represented below:

H 1A: The highest UPFS will occur in a high importance function internally attributed.

H 1B: The lowest UPFS will occur in a low importance function externally attributed.

H 1C: The UPFS of a low importance function internally attributed is higher than the UPFS of a high importance function externally attributed, respecting *H 1A* and *H 1B*.

Based on a assumption check it is justified to use ANOVA, see appendix 11G for a detailed discussion.
For testing the hypotheses 1A, 1B and 1C a comparison between the average ERS-scores of the different scenarios has to be done. Table 11.7 gives the result of this comparison.

Table 11.7: Multiple comparison of ERS-scores between scenarios (p-values)

Scenario*	avg ERS	Scenario*	1	2	3	4
1	19.79	1		0.20	0.83	0.01
2	17.64	2	0.20		0.69	0.67
3	18.85	3	0.83	0.69		0.12
4	16.45	4	0.01	0.70	0.12	

Scenario 1: Failure Importance high, Failure Attribution intern
Scenario 2: Failure Importance high, Failure Attribution extern
Scenario 3: Failure Importance low, Failure Attribution intern
Scenario 4: Failure Importance low, Failure Attribution extern

There is only a significant difference in ERS-score between scenario 1 and scenario 4. So, based on this result it can be concluded that the a failure of a high important function internally caused leads to higher level of UPFS than a low important function externally caused.

Hypothesis 1A: 'the highest UPFS will occur in a high importance function internally attributed" can not be fully supported. Scenario 1 has the highest UPFS but this score is not significantly different from scenario 2 and scenario 3. However the UPFS of scenario 1 differs significantly from the UPFS of

scenario 4. So, there is a strong indication that the highest UPFS will occur in a situation when a high importance function fails, internally attributed.

Hypothesis 1B: "the lowest UPFS will occur in a low importance function externally attributed" can not be fully supported". Scenario 4 has the lowest UPFS but this score is not significantly different from scenario 2 and scenario 3. However the UPFS of scenario 4 differs significantly from the UPFS of scenario 1. So, there is a strong indication that the lowest UPFS will occur in a situation when a low importance function fails externally attributed.

Hypothesis 1C: "the UPFS of a low importance function internally attributed is higher than the UPFS of a high importance function externally attributed, respecting *H 1A* and *H 1B"* can not be fully supported. The UPFS of a low importance function internally attributed is higher than the UPFS of a high importance function externally attributed, (respecting *H 1A* and *H 1B*), see table 11.7, but not significantly. These results and indications can be presented in the figures as given below:

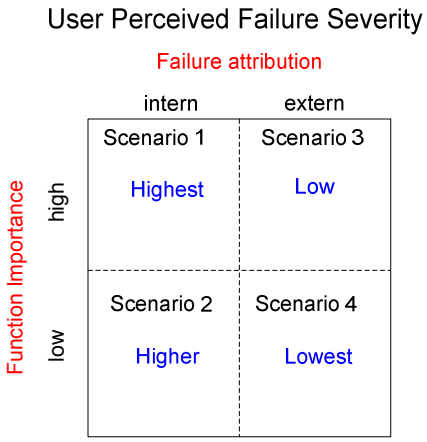


Figure 11.2: UPFS-matrix: relationship between function importance and failure attribution

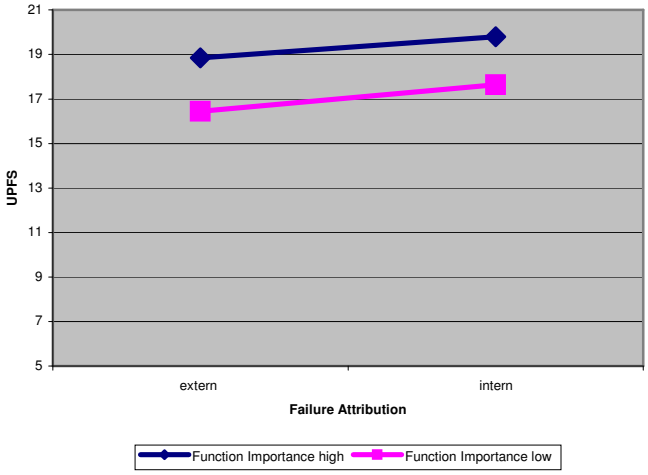


Figure 11.3: Relationship between function importance, failure attribution and UPFS

The results above indicate that both Function Importance and Failure Attribution influence UPFS. The UPFS increases when failure attribution moves from extern to intern. A failure in a high importance function internally (externally) attributed leads to higher level of UPFS than a failure in a low importance function internally (externally) attributed. Therefore a two way ANOVA is conducted.

Table 11.8: Result two-way ANOVA dependent variable: ERS-score 1

	p-value
Function Importance	0.16
Failure Attribution	< 0.01

Adj. $R^2 = 0.08$

Despite the results of the qualitative analysis, as presented in figure 11.4, the variable Function Importance is not significant. The variable Failure Attribution is significant. Since the two lines of Function Importance high and Function Importance low in figure 11.4 are nearly parallel it is not expected that there is any significant interaction between these two variables. ANOVA confirms this, see table 11.9.

Table 11.9: Result two-way ANOVA dependent variable: ERS-score 2

	p-value
Function Importance	0.16
Failure Attribution	< 0.01
Function Importance * Failure Attribution	0.87

Adj. R^2=0.07

In the first experiment there was a significant difference in UPFS based on Function Importance. However in this experiment Function Importance is not significant. Only failure Attribution influences UPFS. So the difference in UPFS-score could be explained by only Failure Attribution. In other words the UPFS-score of scenario 1 and 2 combined should be significant different from the UPFS-score of scenario 3 and 4 combined. ANOVA reveals that FA is highly significant ($p<0.01$). It is also interesting to test whether Function Importance does not have any influence on UPFS when Failure Attribution is not taking into account. ANOVA will be used for this purpose. Scenario 1 and 3 are combined and scenario 2 and 4 are combined. The UPFS-scores do not differ on Function Importance ($p = 0.14$). This result is different with regard to the result of experiment 1. A difference between experiment 1 and 2 is that experiment is conducted in a laboratory setting and experiment 2 in public places for students. This could maybe explain why there is a difference in significance. From table 11.10 it is visible that the difference in ERS-score between the motorized swivel and picture scenario is larger in experiment 1 than in experiment 2 (2.58 vs. 1.17). The ERS-score of the motorized swivel scenario in experiment 1 is significantly higher than the ERS-score of the motorized swivel scenario in experiment 2 ($p < 0.01$). The ERS-scores of the picture scenario in experiment 1 is significantly higher than the ERS-score of the picture scenario in experiment 2 ($p < 0.01$).

In experiment 2 are the difference in ERS-score between the picture and motorized swivel scenario not large enough to detect significant differences. The laboratory setting of experiment 1 approximated more the real-live situations than the setting of experiment 2. In the first experiment participants were confronted with/perceived a failure. In the second experiment participants evaluated a failure. The emotional reactions in the second experiment are therefore less extensive. In real-life it is expected that the emotional reactions on a failure are even heavier than in the first experiment.

Table 11.10: ERS-scores and FI-scores of experiment 1 and 2

	Experiment 1 (N=24)				Experiment 2 (N=113)			
	FI-score		ERS-score		FI-score		ERS-score	
scenario	Mean	SD	Mean	SD	Mean	SD	Mean	SD
motorized swivel	3.75	1.76	19.50	3.26	2.75	1.44	17.54	4.27
picture	9.33	0.88	22.08	3.06	9.38	0.68	18.71	4.02
difference	5.58		2.58		6.63		1.17	

Another difference is that in the first experiment the actual Function Importance scores[3] are used instead of low or high importance. Doing ANOVA with the actual Function Importance scores lead to the following result see table 11.11. In this case is besides the main effect of Failure Attribution also the interaction effect between Function Importance and Failure Attribution significant.

[3] The actual importance scores are the scores filled in by the participant in Function Importance Questionnaire see figure 7.3 and figure 10.1

Table 11.11: Result ANOVA dependent variable: ERS-score

	p-value
Function Importance	0.93
Failure Attribution	< 0.01
Function Importance * Failure Attribution	0.01

Adj. R^2=0.16

However, no clear that pattern is visible between the actual Function Importance, Failure Attribution and UPFS see appendix 11H. Regression analysis confirms that there is no significant interaction effect between the actual Function Importance and Failure Attribution, table 11.12. The significance of actual Function Importance is 0.17. As noted earlier the difference in ERS score between the scenarios is in the second experiment too small to detect differences between the Function Importance levels. An experiment where the participants are confronted with/perceive a failure instead of evaluate a failure would detect significant differences. Therefore Function Importance is significant in experiment 1.

Table 11.12: Regression analysis dependent variable: ERS-score

	p-value
Function Importance	0.17
Failure Attribution	< 0.01
Function Importance * Failure Attribution	0.11

Adj. R^2=0.16

Based on the results of these experiments the research question presented below can be answered.

"To what extent do Failure Attribution (FA) and Function Importance (FI) influence UPFS?"

Based on the results above it has been proven that when the Failure Attribution moves from extern to intern the UPFS-level increases. From the first experiment is concluded that a failure in a high important function causes more irritation than a failure in a low important function. Based on the results of experiment two can be concluded that the effect of Failure Attribution is more important than the effect of Function Importance. Because based on experiment 2, there is a strong indication that a failure of a high important function internally attributed leads to a higher UPFS-level than a failure of a low important function internally attributed. Also the same indication is true for a failure externally attributed: a failure of a high important function externally attributed leads to a higher UPFS-level than a failure of a low important function externally attributed leads to a higher UPFS-level. These effects can be represented as follows, figure 11.6.

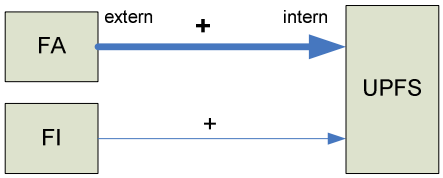


Figure 11.4: UPFS-model relationship between FA and FI

It is also interesting to examine the correlation matrix, see appendix 11I. The Failure Attribution Questionnaire-score (FAQ-Score) is significantly correlated with attribution. This result is obvious. The FAQ-score measures whether the participants perceived their failure attribution scenario as intended. So, a significant large correlation is expected. The magnitude of the correlation is 0.56 highly significant ($p<0.00$). The magnitude of the correlation could be larger when taking into account that the design of scenario 4 could be further improved. Since, the correlation between Failure Attribution and ERS-score

and the correlation between Failure Attribution and FAQ-score is significant, is it not strange that the correlation between FAQ-score and ERS-score is also significant.

Secondly, Age does not have any influences on ERS-score. This result is expected since a homogenous group of students is used. These students have nearly all the same age (average age = 22 years, standard deviation = 2.6 years).

After controlling for Irritableness, Negative Affect does not have any effect on ERS-score anymore. This result proofs that successfully is controlled for Irritableness.

Actual Function Importance is not significantly correlated with ERS-score this could be expected since regression analysis revealed that actual FI is not significantly related with ERS, see also appendix 11H.

A remarkable point is the significant correlation between the Failure Impact Questionnaire score (FIQ-score) and ERS-score. In the case of perfectly controlled for Failure Impact, the score on the FIQ should be the same for all participants (in all scenarios). Table 11.5 shows that there is no significant difference in all four scenarios for the average failure impact. Consequently it could be expected that this on average constant FIQ-score should have no effect on the ERS-score. Apparently this is not the case, since there is a significant correlation ($p<0.00$) of .32. An increase in perceived Failure Impact will lead to a significant increase in ERS-score, see also appendix 11J.

To derive a final model a regression analysis is done with FI, FA, FIQ, NA, Age, FI*FA, FI*FIQ and FA*FIQ as independent variables. The end-model contains the variables FIQ and the interaction-effect of FA and FIQ. Doing forward and backward regression with the same variables led to the same result as stepwise regression. This can be visualized by figure 11.5. It has to be noted that the main effect will not be interpreted when the interaction-effect is significant.

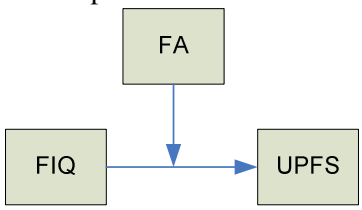


Figure 11.5: UPFS model 1 based on regression

From stepwise it becomes clear that FA, FA*FIQ and FA*FI are both candidates for being included in the model in step 2 (see appendix 11K). In step 2 FA*FIQ is added. The Variance Inflation Factor (VIF) of FA increases enormously. VIF is the reciprocal of tolerance. Tolerance is the proportion of a variable's variance not accounted for by other independent variables of the model. So, a variable with a high VIF value contributes little information to a model. Large VIF values are an indicator of multicollinearity. This means that FA and FA*FIQ almost explain the same variance in ERS-score. So, it does not matter that much which of the two variables FA or FA*FIQ is added. Doing this regression analysis again but in this case with two blocks will lead to different results see Appendix 11L. The first block contains: Age, NA, FI, FA and FIQ; the second block contains FI*FA, FI*FIQ en FA*FIQ. Stepwise regression will first examine the variables in block 1 and will than continue with block 2.

With this regression method, choosing first from the main effects (block 1) and than choosing from the interaction effects (block 2) leads to another model, figure 11.6.

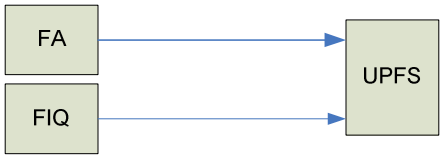


Figure 11.6: UPFS model 2 based on regression

From appendix 11L is visible that when FA is added to the model the VIF of the FA*FIQ rises heavily. This confirms the suggestion that it does not matter which factor is added: FA or FIQ*FA. When one of the two factors is added the other becomes redundant. In means of Adj. R^2 both models have the same value of 0.20. So, both models explain the same percentage of total variation of ERS-score. Based on these results it is clear that FA and FIQ have a significant effect on ERS-score but whether this is an interaction effect or just a main effect is disputable. However for both models applies: When the Failure Impact increases, UPFS increases. When the failure attribution moves from extern to intern, the UPFS increases.

Doing stepwise regression analysis with the actual Function Importance leads to the same models as given in figure 11.5 and figure 11.6. As noted earlier the first 35 participants get a different Failure Impact Questionnaire. Deleting these participants and doing regression analysis again (stepwise and stepwise in blocks) led to the same result, namely figure 11.8 and figure 11.9.

In conclusion, H1A and H1B are not fully supported but there is a strong indication that a failure of a high importance function internally attributed has the highest UPFS and that a failure of a low importance function externally attributed has the lowest UPFS. There is a significant difference between a failure in a high importance function internally attributed and a failure of a low function importance externally attributed, see also figure 11.4.

H1C is not fully supported but there is a strong indication that the UPFS of a low importance function internally attributed is the same as the UPFS of a high importance function externally attributed, respecting *H 1A* and *H 1B,* see also figure 11.3 .

According to the results of the regression analysis it can to be concluded that Failure Attribution and (Perceived) Failure Impact have a significant influence on UPFS. When the failure attribution moves from extern to intern, UPFS increases. When the failure impact increases, UPFS increases. It is unclear whether Failure Attribution and Failure Impact influence UPFS through an interaction effect or just by their main effects. For easiness of interpretation the assumption is made that UPFS will be influenced through the main effects of Failure Attribution and Failure Impact. From the first experiment can be concluded that Function Importance significantly influences the UPFS-level. Failures in high important functions lead to a higher level of UPFS. The results of both experiments are combined to one model figure 11.7. Based on the Beta's Failure Attribution and Function Importance have nearly the same effect. Function Importance has a much smaller effect. Therefore the UPFS model is presented as in figure 11.7.

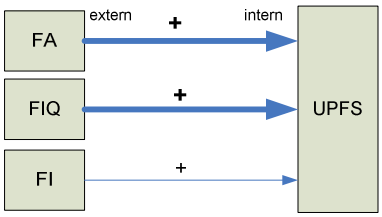


Figure 11.7: UPFS end-model

12. Conclusions and recommendations

This chapter will summarize the answers on the research questions and some other relevant results will be discussed. The implications of this thesis will be discussed and some recommendations will be made for further research.

12.1 Conclusions

The research is conducted at the Eindhoven University of Technology. Participants were all students. Generalization of these results has to be done with care. Especially, when these results are generalized across mature users (>30 years). Another limitation is that for the experiment a high-end consumer electronic product, worth € 3000, is used. Generalization to products which are much cheaper may be not effective.

Two experiments were conducted for measuring User Perceived Failure Severity (UPFS). UPFS is the level of user-irritation caused by a product failure. Irritation is an emotion (reaction). Emotion reactions consist of stable traits and fluctuating affective transitory states. Stable traits describe how people usually or typically are. Affective states describe the affective condition of people during a specific moment in time. Affect refers to the feeling side of consciousness, like pleasure and displeasure, happiness and sadness, liking and disliking. Irritation is also an emotion.

There needs to be controlled for "irritableness" in consumer research that measure irritation caused by a product failure. Irritableness can be seen as how quickly people get irritated. This level of irritableness determines the irritation reaction of users confronted with a failure in a consumer electronic product. Users may overreact or "underreact". Therefore users need to be selected with an average normal level of irritableness. A scale that could measure irritableness has to fulfill certain criteria. These criteria are that the scale needs to measure the personal trait: irritableness. The scale has to be valid and reliable. The scale should be easy enough to use for the participant and the researcher. Based on these criteria, PANAS(-X) is suitable for measuring irritableness. PANAS-X is the expanded form of PANAS. PANAS is the Positive and Negative Affect Schedule, which can measure emotional experience. For this application PANAS-(X) was translated into Dutch. Based on an analysis of 200 Dutch PANAS-X questionnaires can be concluded that the use of the original PANAS can control for irritableness. The scale Negative Affect of PANAS is a measurement-instrument for irritableness. Based on the results of the two experiments (N=348), Dutch students with a Negative Affect level between 13.43 and 23.95 have a normal irritableness level. Results confirm the validity and reliability of the Dutch version of PANAS.

Also some conclusion can be drawn about the Dutch version of the eXpanded form of PANAS. The reliability of the scales of the Dutch version of PANAS-X are reasonably well. Especially the scales related to Negative and Positive Affect. Two scales (Shyness and Serenity) measuring "other affective" states don't posses the minimal acceptable reliability level. The reliability of the scale Shyness is low and the reliability of Serenity is unacceptable low. With regard to validity can be said that the basic positive emotions scale (Joviality, Self-Assurance and Attentiveness) have sufficient convergent and discriminant validity. The basic negative emotion scales (Fear, Hostility, Guilt and Sadness) have sufficient convergent validity but there is a lack of discriminant validity. This means that it is not possible to specify the Negative Affect Emotions into Fear, Hostility, Guilt, or Sadness.

For measuring UPFS a scale is needed that measures irritation. For measuring UPFS not the stable traits (as were the case for irritableness) are at interest but the affective transitory states. Besides this criterion the measurement instrument needs to be valid and reliable and also easy to use for participants and researchers. Based on the results of the two experiments, the irritation caused by a failure in a consumer

electronic product (UPFS) can be measured with the Emotional Response Scale. This scale is reliable and valid and easy to use. PANAS-X can not measure UPFS. The reason is that PANAS-X can not detect differences between a "normal situation" and a failure situation well. However the Emotional Response Scale can measure UPFS, is valid and reliable and easy enough to use for participants as well as for the researchers.

The result of the first experiment (N=25) is that besides failure importance: the relative importance of the function affected by the failure, also failure attribution plays an important role in UPFS. Failure attribution means whether the cause is internally attributed or externally attributed. Internal attribution is that the cause of a failure is attributed to the participant or the television itself. External are all things beyond internal, like blaming the weather, the cable company or the power company for a failure. The level of UPS for these different scenarios is visualized in figure 12.1. This hypothetical UPFS-matrix is based on quantitative and qualitative results of the first experiment. However, the Emotional Response Scale can measure UPFS, is valid and reliable and easy enough to use for participants as well as for the researchers.

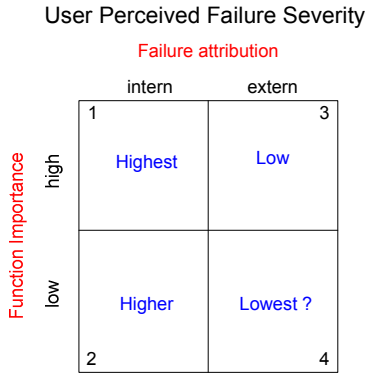


Figure 12.1: UPFS-matrix

In the second experiment the UPFS matrix was tested (N=149). Four different scenarios were tested and the UPFS-levels of these four scenarios were compared. The numbers in the UPFS-matrix represent the different scenarios. The highest UPFS occurs in a situation where a high importance function internally attributed fails (scenario 1). This UPFS is significant different from the situation where a low importance function externally attributed fails (scenario 4), although not significant different from the other scenarios (scenario 2 and 3). The UPFS of a low importance function internally attributed failure (scenario 2) is higher than the UPFS of a high importance function externally attributed failure (scenario 3). However the difference is not significant.

The UPFS-level is influenced significantly by Failure Attribution and perceived Failure Impact. Failure Impact is the loss of functionality as a result of the failure. When the Failure Attribution moves from extern to intern UPFS increases. When the Failure Impact increases the UPFS also increases. Function Importance is not significant in the second experiment but is significant in the first experiment. An explanation could be that in the second experiment participants are asked to evaluate the failures. In the first experiment people are confronted with a failure while doing a task. The emotional reaction in the second experiment is less than in the first experiment. Therefore the differences found in function importance in the second experiment are not significant but are nevertheless in the proposed direction. Based on the results of both experiment the UPFS end-model can be presented, figure 12.2. Thicker lines stand for larger effects on UPFS.

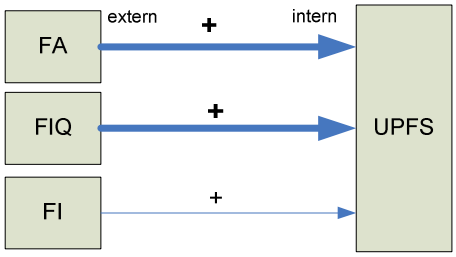


Figure 12.2: UPFS end-model

The obvious results with regard to failure impact imply that a larger failure impact will lead to a higher UPFS-level. More interesting is that despite the fact that all participants were exposed to the same level of failure impact, they perceived this failure impact rather differently and consequently this resulted in a different UPFS-level. It is confirmed that failure attribution influences UPFS (second experiment). It has to be noted that an externally attributed failure which actually is an internal failure will not be perceived externally enduringly. For example, in the situation where there is a failure in the picture you could attribute this failure externally to the cable company. However when this failure occurs several times it is likely that due to social contact (with the neighbors) it becomes clear that your television is the only television which has a problem with the picture. In this situation it is likely that the users will (re)attribute the picture failure internally. From the first experiment it is known that whether the failures are attributed internally or externally is person dependent. Also just as failure impact the results with failure attribution are person specific. Therefore UPFS is largely person dependent. How important a function is also depends on the user. Gielen 2007 concludes that there is moderate agreement between participants on the importance of functions. This is confirmed in this thesis project. For high importance functions there is high agreement on the importance of the function. For low importance functions there is less agreement on the importance of the functions. This can be explained by the fact that less important function are relatively new functions like Ambilight and the motorized swivel. Users need to explore these functions to capture the real added value of these functions and to evaluate these functions if they are important to them or not.

In conclusion, User Perceived Failure Severity is a matter of taste.

High-end consumer electronic manufacturers should take into account that user perceived failure severity is heavily user dependent. For reliability optimization, users with the same characteristics with regard to failure attribution, failure impact and function importance should be categorized. This means that it is not enough to classify users only based on the Rogers-curve (1983) like early adaptors, or laggards. An extension should be made with the factors listed above. Based on this segmentation of users, different market strategies can be developed. Product segmentation means that some functions which are not completely developed and tested, but are from marketing viewpoint very attracting to add, can be included in one segment and excluded in the other segment. Users with a low perceived failure impact level who tend to attribute failures mostly external will perceive a failure less severe than a user with a high perceived failure impact level mostly attributing failures internal. So, for the first group it is attractive from marketing and reliability viewpoint to add this function in the product for this group, since the UPFS will be low when a not completely developed and tested function fails. Based on function importance a prioritization of failures can be made. However, also the importances of new functions are partly user specific. So, resolving a specific failure will add more value to one segment than to the other segment. Therefore the failure of a function that is seen by the most profitable user group as most important should be resolved first. In conclusion this can be summarized in the figure below.

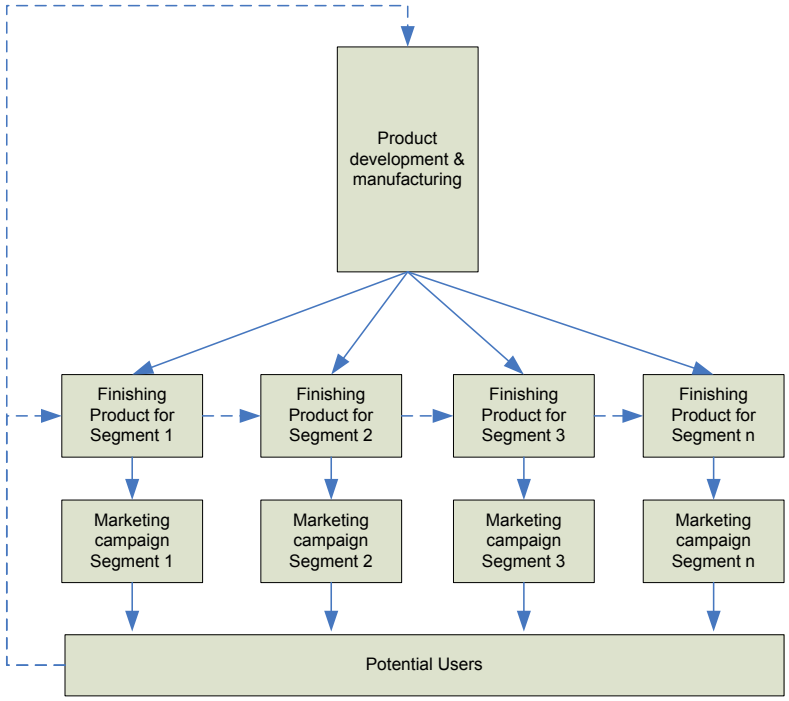


Figure 12.3: Segmentation by use of the UPFS model

Based on (marketing) research (potential) users are categorized in different segment based on Failure Impact, Failure Attribution and Function Importance. The "base-product" will be adjusted to the specific users segments in order to minimize reliability problems. This means that certain functions are included for some segments but are excluded for other segments. Specific marketing campaigns should take care for the distribution of the specific segmented products to the accompanying (segmented) users.

12.2 Recommendations for further research

More research needs to be done whether Dutch students do not have specific emotions like Fear, Hostility, Guilt or Sadness but have a more general emotion reaction. In other words when Dutch students have an emotion reaction like Fear does this also simulates emotions with regard to Hostility, Guilt and Sadness? Is it therefore not possible to distinguish between Fear, Hostility, Guilt or Sadness in this thesis? Another option is that some negative affect items which are used in PANAS-X (and not in PANAS) are not translated well. PANAS-X is developed and validated largely in the USA. PANAS-X is translated into Dutch for use in the Netherlands. In these cultural different countries the same words and phrases could mean different things and so describe different emotions.

For reliability optimization more research has to be done on user characteristics. In that research project should be investigated how (and why) users perceive the impact of the same failure differently and when they attribute the failure internally or externally. With this knowledge user groups could be defined and for every group a specific reliability optimization process could be developed.

Failure Impact, Failure Attribution and (Function Importance) explain 20% of the variance in the UPFS-score. Also other failure characteristics could influence UPFS. As indicated by De Visser (2006) there are also other relevant characteristics like Failure Frequency, Failure Reproducibility, Failure Solvability and Failure Work Around. Research taking into account these failure characteristics will lead to a model which can explain and predict UPFS better.

References

Allport GW. *Personality: a Psychological Interpretation*. Holt, Rinehart & Winston; New York; 1937.

Anderson CA, Lindsay JJ, Bushman BJ. "Research in psychological laboratory truth or triviality?". *Current directions in psychological science* 1999; 8; p. 3-9.

Averill JR. *Anger and Aggression: An essay on emotion*. Springer: New York; 1982.

Averill JR. "Studies on anger and aggression." *American Psychologist* 1983; 38; p. 1145-1160.

Bagozzi RP. "An examination of the Psychometric properties of measures of negative affect on the PANAS-X scales". *Journal of Personality and Social Psychology* 1993; 65; 4, p. 836-851.

Bandura A. *Social Foundation of Thought and Action: A social Cognitive Theory.* Englewood Cliffs, NJ: Prentice Hall; 1986.

Baron RA. *Human Agression*, NY Plenum: New York; 1977.

Berden TPJ, Brombacher AC, Sander PC. "The building bricks of product quality: an overview of some basic concepts and principles". *International Journal of Product Economics* 2000; 67; p. 3-15.

Borenstein M. 2001: Power and Precision 2.1, May 15 2007, http://www.power-analysis.com/power_analysis.htm

Brooks B, Schofield N. "Time-to-market: time equals money –where does it all go?". *World Class Design to Manufacture* 1995; 2; 6; p. 4-10

Brombacher AC, de Graef MR. *Anticiperen op trends, Betrouwbaarheid van technische systemen.*, Edited by Dr. de Graef MR, Stichting Toekomstbeeld der Techniek: Den Haag; 2001.

Brombacher AC. Trends in the reliability analysis of consumer electronics products. International publication Philips Institute of Industrial Support (CFT); 2002

Brombacher AC, Sander PC, Sonnemans PJM, Rouvroye JL. "Managing product reliability in business processes "under pressure"". *Reliability engineering and system safety* 2005; 88; p. 137-146.

Chaplin WF, John OP, Goldberg LR. "Conceptions of states and traits dimensional attributes with ideas as prototypes". *Journal of Personality and Social Psychology* 1988; 54; p. 541-557.

Clark LA, Watson D. "Mood and the mundane: Relations between daily life events and self-reported mood". *Journal of Personality and Social Psychology* 1988; 54; p. 296-308.

Clark LA, Watson D, Leeka J. "Diurnal variation in the positive affects". *Motivation and Emotion* 1989; 13; p. 205-234.

Classen A, Lopez LA: "New product introduction between two geographically dispersed entities". *International Conference on Pioneering New Technologies: Management Issues and Challenges in the Third Millennium* 1998; p. 535-540.

Conway JM, Huffcutt AI . “A review and evaluation of exploratory factor analysis practices in organizational research”. *Organizational Methods* 2003; 6; 2; p. 147-168.

Davidz JR. *The language of emotion*. Academic Press: New York 1969.

Dijkstra L, Dirne CWGM, Govers CPM, Sander PC. *Samenwerking in Ontwikkeling*. Kluwer: Deventer; 1997

Van Dolen W, Lemmink J, Mattson J. Rhoen I. “Affective consumer responses in service encounters: The emotional content in narratives of critical incidents”. *Journal of Economic Psychology* 2001; 22; p. 359-376

Dormann C, Zapf D. “Social stressors at work, irritation, and depressive symptoms: Accounting for unmeasured third variables in a multi-wave study”. *Journal of Occupational Psychology* 2002; 75; p. 33-38.

Endler NS, Mangusson D. “The international model of anxiety: An empirical test in an examination situation”. *Canadian Journal of Behavioural Science* 1976; 9; p. 101-107.

Fennis BM, Bakker AB. “Stay tuned- We will be back right after these messages”: Need to evaluate moderates the transfer of irritation in advertising”. *Journal of Advertising* 2001; 30; 3; p. 15-25.

Foo SW, Lien WL, Xie W, Geest van E. “Reliability by design: A tool to reduce time to market”, *IEEE Engineering Management Conference* 1995; p. 251-256 .

Fowles DC. “Application of behavioral theory of motivation to the concept of anxiety and impulsivity”. *Journal of research in Personality* 1987; 21; p. 417-435.

Gencoz T. “Positive and negative affect schedule: a study of validity and reliability”. *Turk Piskoloji Dergisi* 2000; 15; p. 19-28.

Gibson JL, Ivancevich JM, Donnelly JH, Konopaske R. *Organizations, Behavior Structure Processes.* McGraw-Hill Irwin: New York; 2003.

Gielen SWF. The influence of Function Importance on User Perceived Failure Severity. *Master Thesis*, Eindhoven University of Technology, 2007.

Goodman JH. “It might not be your product” *Quality progress* 2002; 35; 4; p. 73- 78.

Goodwin CJ. *Research in psychology: methods and design*. John Wiley & Sons Inc.: Hoboken; 2005.

Gray JA. “Perspective on anxiety and impulsivity: A commentary”. *Journal of Research in Personality* 1987; 21; p. 493-509.

Green SB, Salkind NJ. *Using SPSS for Windows and Macintosh*. Pearson Education Inc.: Upper Saddle River, New Jersey; 2002.

Hansen R, “Commonsense Attribution”, *Journal of Personality and Social Psychology*; December 1980; 39; p. 996-1009.

Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. *Multivariate Data Analysis.* Pearson Prentice Hall: Upper Saddle River, New Jersey; 2005.

Heider F. *The Psychology of Interpersonal Relations*. Wiley: New York 1958.

Hill RD, van Boxtel MPJ, Ponds R, Houx PJ, Jolles J. "Positive affect and its relationships to free recall memory performance in a sample of older Dutch from the Maastricht Aging Study". *International Journal of Geriatric Psychiatry* 2005; 20; p. 429-435.

Hofmann BU, Meyer TD. "Mood fluctuations in people putatively at risk for bipolar disorders, *British Journal Psychological Society* 2006; 45; p. 105-110.

Hughes GD. "Selecting scales to measure Attitude Change", *Journal of Marketing research* 1967; 4; p. 85-87.

Ihrig A, Hoffmann J, Triebig G. "Examination of the influence of personal traits and habituation on the reporting of complaints at experimental exposure to ammonia" *International Archives of Occupational and Environmental Health* 2006; 79; p. 332-338.

Interbrand Press release, *Best Global Brands*, Amsterdam 27 July 2006

Joiner TE, Sandin B, Chorot P, Lostao L, Marquina G. "Development and factor analytic validation of the SPANAS among women in Spain: (More) cross-cultural convergence in structure of mood". *Journal of Personality Assessment* 1997; 68; p. 600-615.

Jones EE, Michela JL 1972. "The actor and observer : divergent perceptions of the causes of behavior". In: Jones E.E., Kanouse D.E. Kelly H.H., Nisbett R.E., Vallins S., Weiner B. (Eds.), *Attribution: perceiving the causes of behavior*, General Learning Press; Morristown, NJ; 1972; p. 79-94.

Jarvis WGB, Petty RE. "The Need to Evaluate". *Journal of Personality and Social Psychology* 1996; 70; 1, p. 172-194.

Jones J and Hayes J. "Investigation of the occurrence of: no-faults-found in electronic equipment, IEEE transaction on reliability". *IEEE transactions on reliability* 2001; 50; p. 289-292.

Kleuskens YHJ. Beyond adoption: customer classification as a method to predict use behavior for innovate products, *Master Thesis*, Eindhoven University of Technology, 2007.

Krohe HW, Egloff B, Kohlmann CW, Tausch A. "Investigation with a German version of the positive and negative affect schedule (PANAS)". *Diagnostica* 1996; 42; p. 131-168.

Lazarus RS. *Emotion and adaptation*. Oxford University Press: New York, 1991.

Lazar J. "Workplace user frustration with computers: an exploratory investigation of the causes and severity". *Behaviour & Information Technology* 2006; 25; 3; p. 239-251.

Lee DY. "Distribution channels of Philips domestic appliances and personal care products in Chinese economic transition- A case study". *IMP group*, presented at the 17th IMP-conference in Norway; 2001.

Lewis EE. *Introduction to reliability engineering*. John Wiley & Sons Inc: New York; 1996

Locke EA. “Motivation through conscious goal setting, *Applied Preventative Psychology* 1996; 5; p.117-124.

Mascolo MF, Fischer KW 1995. “Developmental transformations in appraisals for pride, shame, and guilt”. In: Tangney JP, Fischer KW (Eds), *Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride,* Guilford: New York; p. 64-113.

McAdams DP 1997. “An introduction to the cognitive science of personality and emotion”. In G. Matthews (Eds), *Cognitive science perspectives on personality and emotion,* Elsevier: New York; 1997. p. 3-30,

McIntyre CW, Watson D, Cunnigham AC. “The effect of social interaction, exercise, and test stress on positive and negative affect”. *Journal of Personality and Social Psychology* 1990, 52, p. 81-90.

Mesquita B, Frijda NH. “Cultural variations in Emotions: A Review”. *Psychological Bulletin* 1992; 112; 2; p. 179-204.

Minderhoud S. “Quality and reliability approach- extending the traditional approach” *Quality and Reliability Engineering International* 1999; 15; 6; p. 417-425.

Mischel W, Shoda Y. “Reconciling processing dynamics and personality dispositions”. *Annual Review of. Psychology* 1998; 48; p. 229-258.

Mohr G. *Die Erfassung psychischer Befindensbeeinträchtigungen bei IndustieArbeitern.*P. Lang: Frankfurt am.Mein; 1986.

Mohr G, Müller A, Rigotti T. “Normwerte der Skala Irritation: Zwei Dimensionen psycherischer Beanspruchung” / “Standardisation data of the Irritation Scale. Two dimensions of mental Strain”. *Diagnostica* 2005; 51; 1; p. 12-20.

Mohr G, Rigotti T, Müller. “A: Irritation- ein Instrument zur Erfassung psychischer Beanspruchung im Arbeitskontext. Skalen- und Itemparameter aus 15 Studien” / “Irritation – an instrument assessing mentak strain in working context. Scale and item parameters from 15 studies”. *Zeitschrift für Arbeits- und Organisationspsychologie* 2005; 49; 1; p. 44-48.

Murthy RKS, Kadur R, Nagaraju N. “Strategic business management in a competitive environment, Engineering Management conference Management in transitions: Engineering a changing world”. *Proceedings of the 1994 IEEE International*; p. 330-337.

Nielsen J. *Usability Engineering.* Academic Press, Inc 538: San Diego; 1993.

Oliver LR. *Satisfaction, A behavioral perspective on the consumer.* Irwin/McGraw-Hill: New York; 1997.

Ouden den E. Development of a Design Analysis Model for Consumer Complaints, Revealing a New Class of Quality Failures. *Dissertation;* Eindhoven University of Technology; 2006.

Patten ML. *Questionnaire research: A practical guide*. Pryczak: Los Angeles; 1998.

Pelsmacker de P, Van den Bergh J. “Advertising content and irritation: A study of 226 TV commercials”. *Journal of International Consumer Marketing* 1998; 10; 4; p. 5-26.

Petkova VT. An analysis of field feedback in consumer electronics industry, *Dissertation*, Eindhoven University of Technology, 2003.

Peterson RA. "On the Use of College Students in Social Science Research: Insights From a Second-Order Meta-Analysis". *Journal of Consumer Research,* 2001; 28; p. 450-461.

Rogers EM. *Diffusions of Innovations*, New York; Free Press.

Roseman IJ 2001. "A model of appraisal in the emotion system: Integrating theory, research, and applications". In: Scherer KR, Schorr A, Johnstone T (Eds.), *Appraisal processes in emotion: Theory, methods, research,* Oxford University Press: New York; 2001; pp. 68-91.

Russell A. "circumplex model of affect". *Journal of Personality and Social Psychology* 1980; 39; p. 1161-1178.

Scherer KR 2001. "Appraisal considered as a process of multi-level sequential checking". In: Scherer KR, Schorr A, Johnstone T (Eds.), *Appraisal processes in emotion: Theory, methods, research,* Oxford University Press: New York; 2001; pp. 92-120.

Scherer KR 1984. "On the nature and function of emotion: A component process approach". In Scherer KR, Elkman P (Eds.), *Cognitive perspectives on emotion and motivation* Hillsdale: NY, Erlbaum; 1984; p. 293-317.

Scully JK. "The hidden crisis in test effectiveness", *Automatic test conference proceedings IEEE systems readiness technology conference* 1998; p. 59-66.

Shaver P, Schwartz J, Krison D, O'Conner C. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology* 1987; 52; p. 1061-1086.

Smith CA, Lazarus RS. "Appraisal components, core relational themes, and the emotions". *Cognition and Emotion* 1993; *7*; p. 233-269.

Smith NC, Thomas RJ, Quelch JA. "A strategic approach to managing product recalls". *Harvard Business Review* 1996; 71; p. 102-112.

Smuckle SC, Egloff B, Burns LR. "The relationship between positive and negative affect in the Positive and Negative Affect Schedule". *Journal of Research in Personality* 2002; 36; 463-475.

Spielberger CD. *Anxiety: Current trends in theory and research.* Academic Press: New York; 1972.

Taylor SE. "The interface of cognitive and social Psychology". In: Harvey JH. (Ed), *Congnition, social behavior, and the environment*, Lawrence Erlbaum Associates, Hillsdale NJ, 1981, p. 297-332.

Terracciano A, McCrea RR, Costa PTJ. "Factorial and construct validity of the Italian positive and Negative affect schedule (PANAS)", *European Journal of Psychological Assessment* 2003; 19; 2; p. 131-141

De Visser IM. "User Centered Reliability Analysis of Consumer Electronics Products". *Dissertation.* version May 2006, Eindhoven University of Technology.

Watson D, Tellegen A. “Toward a Consensual Structure of Mood”. *Psychological Bulletin* 1985; 98; 2; p. 219-235.

Watson D, Clark LA, Tellegen A. “Development and Validation of Brief Measures of Positive and Negative Affect”. *Journal of Personality and Social Psychology* 1988; 54; 6; p. 1063-1070.

Watson D. Intraindividual and interindividual analyses of Positive and Negative Affect: Their relation to health complaints, perceived stress, and daily activities. *Journal of Personality and Social Psychology* 1988; 54; p.1020-1030.

Watson D, Clark LA. Affects separable and inseparable: On the hierarchical arrangement of the negative affects, *Journal of Personality and Social Psychology* 1992; 62; p. 489-505.

Watson D, Clark LA. “The PANAS-X, Manual for the Positive and Negative Affect Schedule-Expanded form”. *unpublished* 1999, University of Iowa.

Watson D, Wiese D, Vaidya J, Tellegen A. “The two general activation systems of affect: structural findings, evolutionary considerations, and psychobiological evidence”. *Journal of Personality and Social Psychology* 1999; 76; 5; p. 820-838.

Wheelwright SC, Clark KB. *Revolutionizing product development: Quantum leaps in speed, efficiency and quality* Free Press: New York; 1992.

Weiner B, Russel D, Lerman D. “The cognition-emotion process in achievement-related contexts”. *Journal of Personality and Social Psychology* 1979; 37; 1211-1220.

Weiner B. “An attributional theory of achievement motivation and emotion”, *Psychological Review* 1985a; 92; 4; p. 548-573.

Weiner B. “‘Spontaneous’ Causal Thinking”, *Psychological Bulletin;* 1985b; 97; January; p. 74-84.

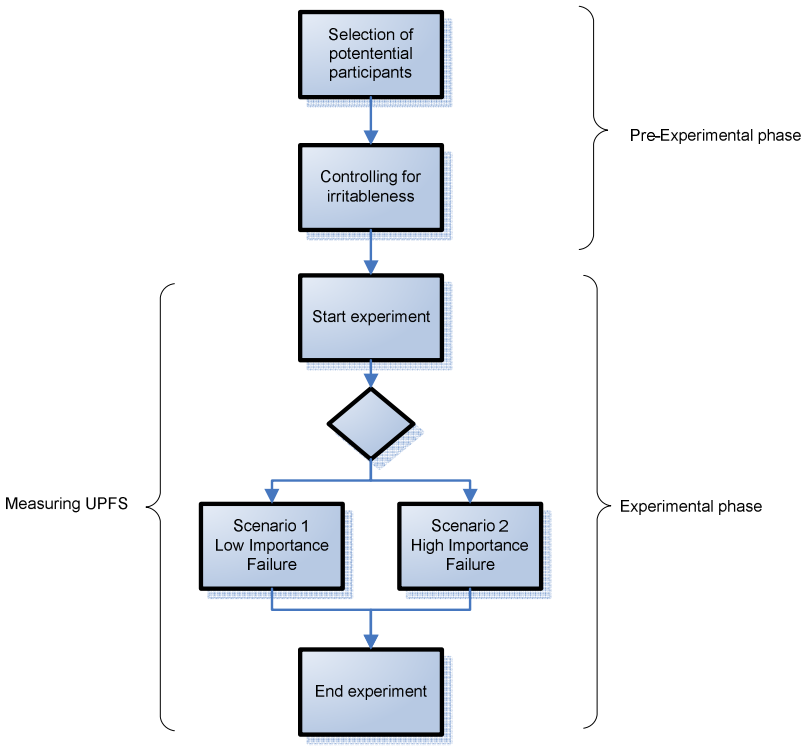
Weiner B. *An attributional theory of motivation and emotion*. Springer: New York; 1986.

Westbrook RA, Oliver RL. “The dimensionality of consumption emotion patterns and customer satisfaction”. *Journal of Consumer Research* 1991; 18; p. 84-91.

Williams R, Banner J, Knowles I, Dube M, Natishan M, Pecht M. “An investigation of 'cannot duplicate' failures”. *Quality and reliability engineering International* 1998; 14; p. 331-337.

Wranik T (2005): “Personality under stress: who gets angry and why? Individual differences in cognitive appraisal and emotion”. *Dissertation.* Theses N° 336, University de Genève

Appendix 2A High level design experiment



Appendix 3A Definitions of relevant emotional terms

Term	Definition
Affect	Affect refers to the feeling side of consciousness. "Feeling includes pleasure and displeasure, happiness and sadness, liking and disliking, and the psychological and visceral sensations brought on by the neural-hormonal bodily systems, such as ecstasy.
Emotion	Emotion includes arousal, various forms of affect, and cognitive interpretations of affect. Emotion is more cognitively involved than affect.
Moods	Moods can be distinguished on the basis of their duration. Mood is a temporary state of pleasant or unpleasant disposition.
(stable) Traits	Stable traits describe how people usually or typically are.
(transitory) States	Transitory states describe the affective condition of people during a specific moment in time.

Source: Oliver (1997) and Davidz (1969)

Appendix 3B PANAS

The PANAS

This scale consists of a number of words that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent [INSERT APPROPRIATE TIME INSTRUCTIONS HERE]. Use the following scale to record your answers.

1	2	3	4	5
very slightly or not at all	a little	moderately	quite a bit	extremely

_____ interested	_____ irritable
_____ distressed	_____ alert
_____ excited	_____ ashamed
_____ upset	_____ inspired
_____ strong	_____ nervous
_____ guilty	_____ determined
_____ scared	_____ attentive
_____ hostile	_____ jittery
_____ enthusiastic	_____ active
_____ proud	_____ afraid

We have used PANAS with the following time instructions:

Moment	(you feel this way right now, that is, at the present moment)
Today	(you have felt this way today)
Past few days	(you have felt this way during the past few days)
Week	(you have felt this way during the past week)
Past few weeks	(you have felt this way during the past few weeks)
Year	(you have felt this way during the past year)
General	(you generally feel this way, that is, how you feel on the average)

Appendix 3C PANAS-(X)

PANAS is developed by Watson, Clark and Tellegen (1988). They developed two 10-item mood scales that comprise Positive Affect (PA) and Negative Affect (NA). Positive Affect (PA) reflects the extent to which a person feels enthusiastic, active, and alert. High PA is a state of high energy, full concentration and pleasurable engagement, whereas low PA is characterized by sadness and lethargy. Negative Affect (NA) is a general dimension of subjective distress and unpleasurable engagement that classifies a range of aversive mood states, including anger contempt, disgust, guilt fear, and nervousness, whereas low NA being a state of calmness and serenity (Watson, Clark and Tellegen, 1988). PANAS is sensible for changes in mood in short time periods, as for measuring traits which are stable in time (Watson et al., 1988). Changes in mood in short times are changes in the transitory state of a person.

PANAS has been used extensively in a variety of studies and has been translated in different languages (Smuckle, Egloff and Burns, 2002) like Russian (Balatsky and Diener, 1993), German (Krohe, Egloff, Kohlmann and Tausch, 1996), Spanish (Joiner et al., 1997), Swedish (Hilleras et al., 1998), Turkish (Gencoz, 2000), Italian (Terracciano et al., 2003) and Dutch (Hill et al. 2005). The experiments revealed that the reliability (see also table 3C1 for definitions of statistical terms) of the PA and NA scales are good; Cronbach's alpha ranged from 0.86 and 0.87 (minimally recommended value 0.7 (Hair, Black, Babin, Anderson and Tatham, 2005)) respectively and the correlation between these scales was quite low with -.09 (Watson et al., 1988). The two factors PA and NA accounted for 62,8% to 68,7% of the common variance. This percentage is used to determine how well a particular factor solution accounts for what all the variables together represent. If all variables are very different from one another this percentage is quite low. There is not a minimal acceptable value for the percentage of common variance. Convergent correlations ranged from .85 to .95 which well above the threshold value of 0.7. Discriminant validity is good because the correlations are quite low ranging from -.02 to -0.18. So, it can be concluded that PANAS is a reliable, valid, and efficient medium for measuring two important moods positive affect and negative affect (Watson et al., 1988).

Table 3C1: Definitions of statistical terms.

Term	Definition
Reliability	Reliability is an assessment of the degree of consistency between multiple measurements of a variable.
Cronbach's alpha	A measure that assesses the consistency of the entire scale.
Common variance	Common variance is defined as that variance in a variable that is shared with all other variables in the analysis.
Validity	Validity is the extent to which a scale or set of measures accurately represents the concept of interest
Convergent correlations (Convergent validity)	Convergent correlations assess the degree to which two measures of the same concept are correlated.
Discriminant validity	Discriminant validity is the degree to which two conceptually similar concepts are distinct. A scale is sufficiently different from other related scales.

Source: Hair et al. 2005

Many PA and NA scales have been developed and studied in a variety of research areas. The two mood factors PA and NA are related to different classes of variables. For example, negative affect is related to poor coping, and frequency of unpleasant events (Watson et al., 1988). PANAS is based on a long line of research into the basis dimensions of affect (Watson, Wiese, Vaidya and Tellegen, 1999). Russell (1980) proposed that mood (affect) consist out of two general basic dimensions, Pleasantness and Activation.

These two dimensions define a circumplex, that is, a model in which mood descriptors can be systematically arranged around the perimeter of a circle.

Figure 3B1 shows a circumplex developed by Watson & Tellegen (1985). In this model terms within the same octant are highly positively correlated. Terms in the adjacent octants are moderately positively correlated. Words 90° apart are in essence unrelated to one another, whereas those 180° apart are opposite in meaning and highly negatively correlated. In this model, in contrast to Russel's model, Negative Affect and Positive Affect are seen as the two basic dimensions. Where Russel advocates that Pleasantness and Activation (or Arousal) are the basic dimensions of mood.

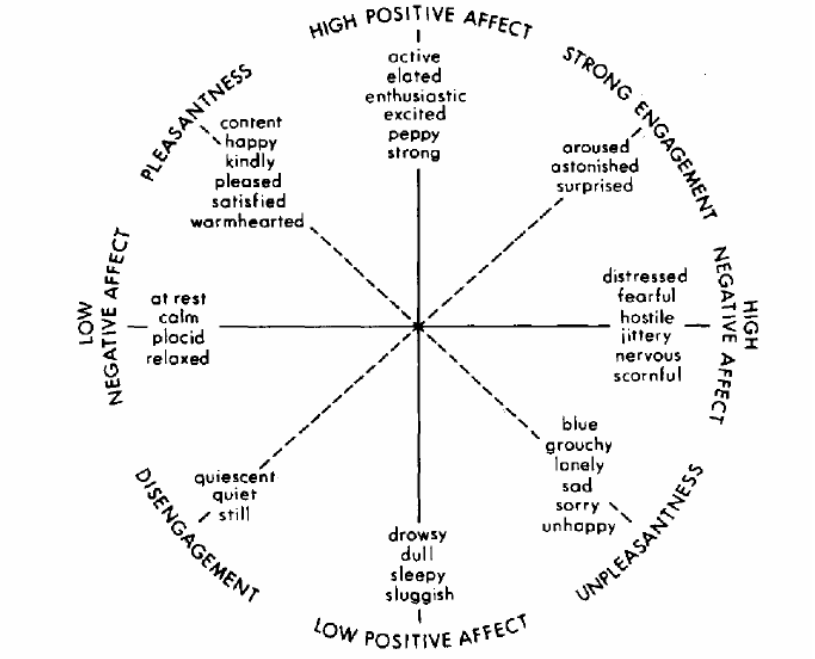


Figure 3C1: Two-dimensional structure of affect (Watson and Tellegen, 1985)

The 20 item PANAS is based on the model above (figure 3B1). It became apparent that this model captures important properties of mood but fails to provide a close fit with the data obtained from the field. (Watson et al., 1999). The Engagement-Disengagement axis deviates from their predicted position in the circumplex (see also figure 3C2).

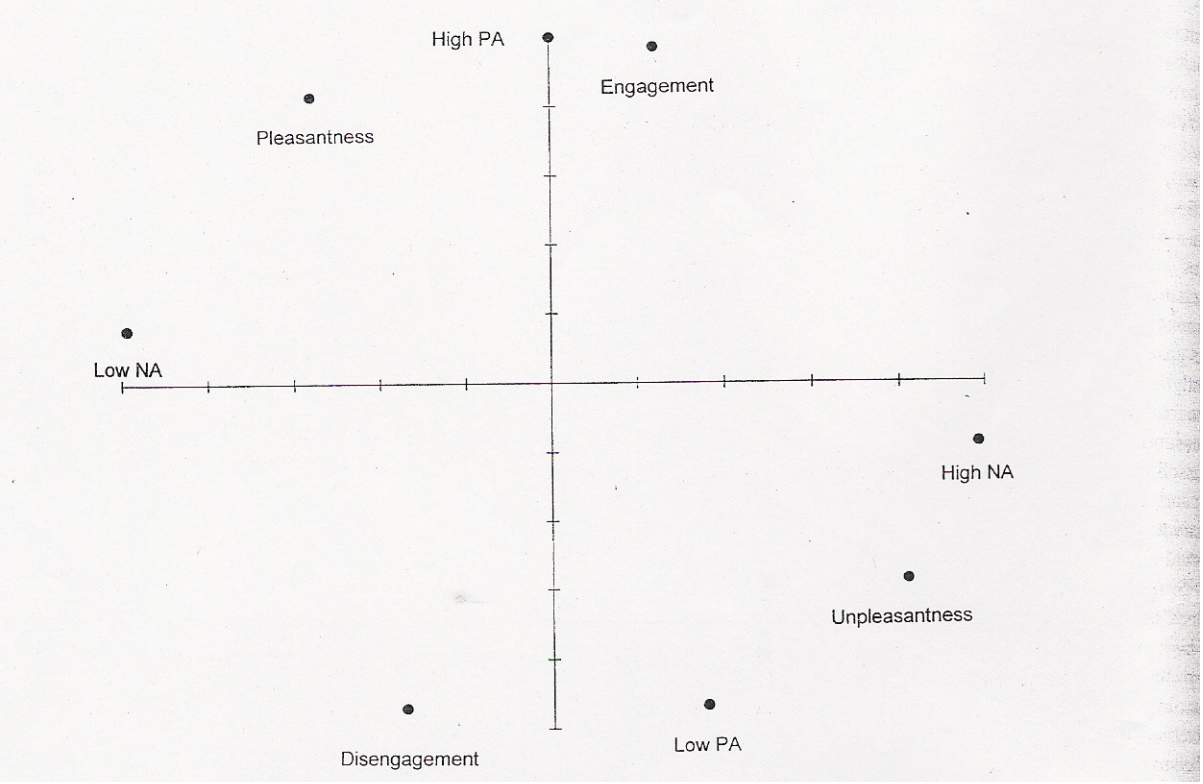


Figure 3C2: factor loadings on the affect circumplex model 1

Most terms cluster in the high positive affect and in the high negative affect octants (Watson and Clark, 1999). Therefore PANAS measures especially the presence of high negative or positive affect and does not measure (well) the dimensions pleasantness and engagement. These results led to some critique on the naming of the scales. To this critique will be returned later on.

Watson and Clark developed PANAS further, trying to resolve this problem. The PANAS eXpanded form (PANAS-X), developed in 1994, was a 60-item questionnaire. These 60 items are one-word items which represent an emotional state like happy, sad, angry etc. For every item participants have to indicate with a score from 1 to 5 to what extent the particular item is applicable to him/her. PANAS-X contains good markers of seven of the eight octants in figure 1.3, only disengagement was not assessed. However there are still problems. The results tend to form an ellipse rather than a circle, see figure 3C3.

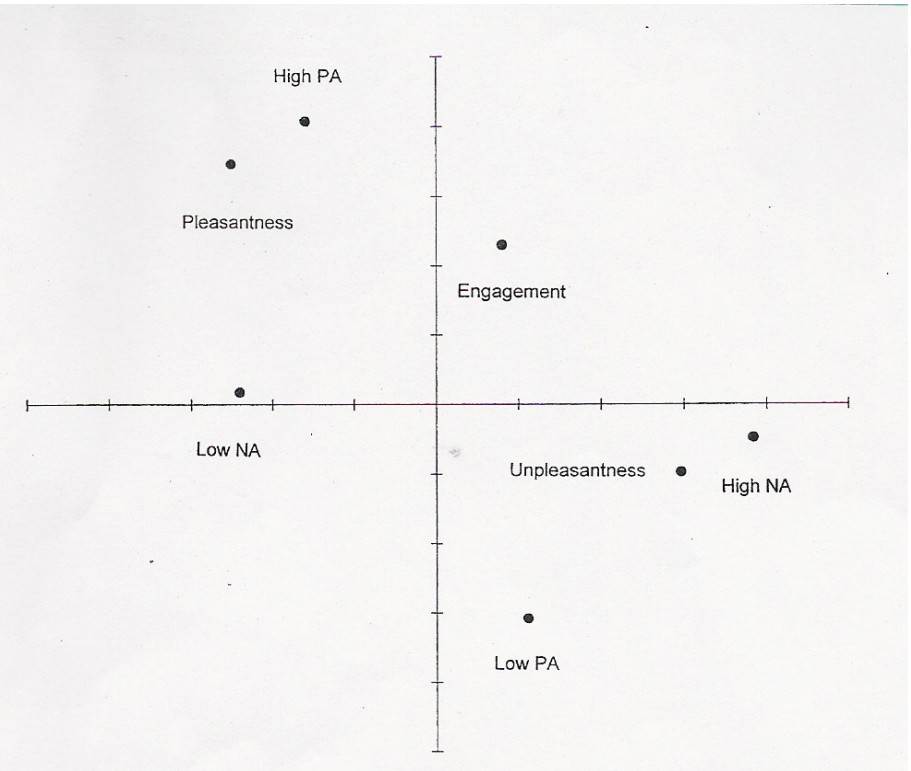


Figure 3C3 factor loadings on the affect circumplex model

This also shows that valance (Pleasantness vs. Unpleasantness) had a much greater influence on mood ratings than did arousel (Engagement vs. Disengagement). In summary, the problems are the unequal spacing of the hypothesized markers and the absence of variables along the Engagement-Disengagement axis. Thus PANAS-X can not measure the whole structure of affect as proposed in figure 3B1. The current PANAS-X misses some markers of the Engagement-Disengagement axis and can not measure this Engagement-Disengagement dimension of affect. In 1999 a new version of PANAS-X was proposed. This PANAS-X consists of two levels. The higher level reflects the "valance" of the mood descriptors (i.e. "whether they represent negative or positive states" (Watson and Clark, 1999). The lower level reflects their specific "content" (i.e. "the distinctive qualities of the individual affects" (Watson and Clark, 1999). These emotional states are measured by a 60-item expanded form of the PANAS (PANAS-X, see figure 3C4).

This scale consists of a number of words and phrases that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent you have felt this way during the past few weeks. Use the following scale to record your answers:

1 very slightly or not at all	2 a little	3 moderately	4 quite a bit	5 extremely
_____ cheerful	_____ sad	_____ active		_____ angry at self
_____ disgusted	_____ calm	_____ guilty		_____ enthusiastic
_____ attentive	_____ afraid	_____ joyful		_____ downhearted
_____ bashful	_____ tired	_____ nervous		_____ sheepish
_____ sluggish	_____ amazed	_____ lonely		_____ distressed
_____ daring	_____ shaky	_____ sleepy		_____ blameworthy
_____ surprised	_____ happy	_____ excited		_____ determined
_____ strong	_____ timid	_____ hostile		_____ frightened
_____ scornful	_____ alone	_____ proud		_____ astonished
_____ relaxed	_____ alert	_____ jittery		_____ interested
_____ irritable	_____ upset	_____ lively		_____ loathing
_____ delighted	_____ angry	_____ ashamed		_____ confident
_____ inspired	_____ bold	_____ at ease		_____ energetic
_____ fearless	_____ blue	_____ scared		_____ concentrating
_____ disgusted with self	_____ shy	_____ drowsy		_____ dissatisfied with self

Figure3C4 PANAS-X: Watson and Clark (1999)

The PANAS-X measures 11 affects (see figure 3C5). Watson and Clark present an enormous amount of data (Watson and Clark, 1999). From these data appears that the PANAS-X scales, are stable over time, show significant convergent and discriminant validity, are highly correlated with corresponding measures of aggregated state affect (short term and long-term), and are strongly and systematically related to measures of personality and emotionality. However one possible exception held for the Surprise-scale when measuring long term state affect. The validity of this scale was not as good as the other scales (Watson and Clark, 1999). Based on these results it is not recommended to use the Surprise-scale.

General Dimension Scales	
Negative Affect (10)	afraid, scared, nervous, jittery, irritable, hostile, guilty, ashamed, upset, distressed
Positive Affect (10)	active, alert, attentive, determined, enthusiastic, excited, inspired, interested, proud, strong
Basic Negative Emotion Scales	
Fear (6)	afraid, scared, frightened, nervous, jittery, shaky
Hostility (6)	angry, hostile, irritable, scornful, disgusted, loathing
Guilt (6)	guilty, ashamed, blameworthy, angry at self, disgusted with self, dissatisfied with self
Sadness (5)	sad, blue, downhearted, alone, lonely
Basic Positive Emotion Scales	
Joviality (8)	happy, joyful, delighted, cheerful, excited, enthusiastic, lively, energetic
Self-Assurance (6)	proud, strong, confident, bold, daring, fearless
Attentiveness (4)	alert, attentive, concentrating, determined
Other Affective States	
Shyness (4)	shy, bashful, sheepish, timid
Fatigue (4)	sleepy, tired, sluggish, drowsy
Serenity (3)	calm, relaxed, at ease
Surprise (3)	amazed, surprised, astonished

Note. The number of terms comprising each scale is shown in parentheses.

Figure 3C5 Item composition of the PANAS-X Scales: Waston and Clark (1999)

According to Watson and Clark PANAS-X scales are sensitive to changing circumstances in a small time period (Moment or Today) (Watson and Clark, 1999). This claim is supported with 4 research studies. The first study among 80 students showed that variations in perceived stress were strongly correlated with fluctuations in Negative Affect (Watson, 1988). In a second study with 196 students had to complete the PANAS-X scales on average seven times a day for 1 week (Clark, Watson and Leeka, 1989). The researchers concluded that Positive Affect varied significant. As hypothesized, Positive Affect rose sharply from early morning until noon, remained quite constant until 9 p.m., and then fell rapidly. Also the results of the first study where replicated: perceived stress is correlated with fluctuations in Negative Affect (Clark and Watson, 1988). In a study conducted by McIntyre, Watson & Cummingham 18 students completed 4 questionnaires within one week. They found evidence that variation in the mood was measured in the PANAS-X scales (McIntyre, Watson and Cunningham, 1990). Positive Affect was increased significantly by social interaction and exercise. Negative Affect was increased significantly by a stressful examination. In a fourth study, 127 students completed two higher order PANAS-X scales (General Positive Affect and Negative Affect) each evening 5 to 7 weeks, 96 of these students also filled in three basic positive emotions scales (Joviality, Self-Assurance, and Attentiveness). All four positive affect scales (the general dimensions scale and the three basic positive emotions were significantly related to social activities) (Watson and Clark, 1992). So, based on these results, the PANAS-X scales are strongly correlated with commonly used measures of state effect and are sensitive to changing conditions (Watson and Clark, 1999). In a recent study conducted by Hofmann and Meyer (2006) PANAS is used for measuring differences, c.q. fluctuations in mood. In this research 89 participants completed daily the two general dimensions scales for 28 days. Based on these results it can be concluded that PANAS-X can be used for measuring person's mood state fluctuations.

There is some critique on the naming of the scales (figure 3B1) as proposed by Watson & Tellegen (1985). One major point of critique is that they misrepresent the actual valence of these dimensions (Watson et al., 1999). Positive Affect implies that this dimension reflects fluctuations in positively valanced mood states (Watson et al., 1999). However (figure 3B1) shows that this dimension also encloses unpleasant negatively valanced terms at its low pole (Watson et al., 1999). Watson et al. 1999 avert this critique with the argument that despite figure 3B1 gives equal weights to both ends (high and low positive of negative affect), the high poles are much more important. A review of the most prominent inventories that assess mood at the affect level indicates that these inventories have much more high-activation terms than low-activation terms (Watson et al., 1999). Another argument is that the terms used to characterize low ends do not define these dimensions themselves (Watson et al., 1999). These low poles reflect the absence of activation rather than the presence of a certain affective state (Watson et al., 1999). So, these dimensions (positive and negative affect) are defined by their activated (high pole) ends. Also the concept of Arousal/Activation can be used for measuring short term affect (ratings of current, momentary mood), but it can not measure long-term differences in emotion (Watson et al., 1999). However, Positive and Negative affect are able to measure affect at both the state and the trait level.

Watson et al. (1999) also present that Positive Affect (PA) and Negative Affect (NA) are components of the biobehavioral systems. These biobehavioral systems consist out of a withdrawal system and an approach system. The withdrawal system (labeled as behavioral inhibition system (BIS)) focuses on maximum attention on analyzing environmental stimuli, especially novel stimuli that could potentially signal danger (Gray, 1987) like a function failure of a television. NA characterizes the subjective component of the withdrawal oriented BIS (Watson et al., 1999). The approach system (behavioral facilitation system (BFS)) stimulates actions that may results in pleasure and reward (Fowles, 1987). Variations in the PA reflect the operation of BFS (Watson et al., 1999). Another remarkable finding is that NA scores remain low in the absence of threat or danger but raise very hard when persons are faced with potential threats (Watson and Clark, 1999), like a function failure of a television. Also remains NA stable over the course of the day. NA can measure a sudden change in state when persons are confronted with an uncomfortable situation like for example a function failure in a television.

Appendix 3D Critical incident question

Consumption emotions refer to the set of emotional responses elicited during product usage or consumption experiences (Westbrook, 1991). Consumption emotion can be conceptualized as distinctive categories of emotional experience and expression (for example joy, anger and fear) (van Dolen at al., 2001). Another approach is to conceptualize consumption emotions as a limited number of structural dimensions underlying emotional categories (for example pleasantness/unpleasantness, relaxation/action, or calmness/excitement (van Dolen at al., 2001). Van Dolen et al. purpose is to develop a classification system to extract emotional content from respondents' written text. A categorical approach appears the most appropriate (van Dolen at al., 2001). They derive emotional content from the respondent's answers to the so-called critical incident question. The critical incident question in their study is a follows: "*Please try to remember a moment, within the last three months, during which you had an extraordinarily positive or negative experience with our company. Can you, in your own words, describe what happened, what has been said, and what you saw. Besides, can you indicate what caused the event*", (van Dolen at al., 2001).

In order to classify the information according to established categories van Dolen et al. used a classification method developed by Shaver, Schwartz, Kirson and O'Conner (1987). This classification is based on the following three levels: the superordinate level (positive vs. negative emotions), the basic level (basic/primary emotions) and the subordinate level (more specific/secondary emotions), (See also figure 3D1). Interrelated reliability ranged from 86% to 100%. Interrater reliability is the extent to which two or more individuals (coders or raters) agree.

They concluded that especially irritation has an extraordinarily negative impact on customer satisfaction (van Dolen at al., 2001). User perceived failure severity tries to capture negative impact on customer satisfaction (caused by failure). That is why the measurement of irritation is so important in the research model of UPFS. This incident analysis can be used by selecting some descriptions of critical incidents, which will be presented to the respondents. Accordingly they have to give a written reaction on these incidents.

Examples of verbalisations classified according to: the super-ordinate, basic and subordinate level and object of emotion

Examples[a]	Super-ordinate	Basic	Sub-ordinate	Object
'You delivered the paper too early in the morning, at a time nobody was in yet. We had an agreement that the paper would be delivered beyond the front door, yet it was standing outside in the rain. I called you and explained the problem. Your company in turn called the company that was responsible for the transportation, while keeping me informed. They asked me whether there was any damage, and pressed me never to carry the paper myself again. After one week, a bouquet of flowers was delivered *to my big surprise*, and you apologised for the inconvenience. That's first rate'!	Positive	Positive surprise	Strong	Organisation
'One morning, one of our copiers broke down. I called your company and, *surprisingly*, the service technician arrived the same day'	Positive	Positive surprise	Weak	Organisation
'When I contacted your company concerning the increasing number of copies, I had a *very pleasurable* experience with your employee who provided me with clear and useful information in this respect'	Positive	Pleasure	Strong	Service employee
'Last February two gentlemen of your company delivered the new copier we ordered. All I can say is that they were *nice and friendly* to me. . .. A *pleasant* encounter'	Positive	Pleasure	Weak	Service employee
'We were *very content* with the service and explanation we received when the one of your employees delivered the 3045 type of copier'	Positive	Contentment	Strong	Service employee
'Your technicians generally manage to solve our problems in a *satisfactory* manner'	Positive	Contentment	Weak	Service employee
'In case of high usage, the paper gets stuck. This is *very annoying*, especially when it happens at night'!!	Negative	Irritation	Strong	Office equipment
'It was *slightly annoying* that the newly arrived copier broke down the same day'	Negative	Irritation	Weak	Office equipment

Examples[a]	Super-ordinate	Basic	Sub-ordinate	Object
'Promise of a color-copier was created by your company. However it turns out that it will not be available in the market before the end of 1997. This is a *big disappointment*'	Negative	Disappointment	Strong	Organisation
'*It is a pity* that moving the copier within our company, from the first to the ground floor should cost so much'	Negative	Disappointment	Weak	Organisation

[a] The examples are all idiomatically translated from Dutch texts.

Figure 3D1: Critical Incident Questions: Van Dolen, Lemmink, Matson, Rhoen (2001)

Appendix 3E Irritation in commercials

Studies on irritation in advertising have primarily focused on attributes that produce negative reactions to that commercial and the brand advertised (Fennis and Bakker, 2001). Fennis and Bakker (2001) extend this line of research by examining the carryover effects of irritation. Research has shown that there are stable individual differences in the tendency to engage in evaluative responding to products, social issues, future behaviors and, so on (Jarvis and Petty, 1996). A Need to Evaluate Scale (NES) is developed and tested (Jarvis and Petty, 1996). In their study people were asked to write about the events of their previous day. People with a high need to evaluate (NE) wrote as twice as many evaluative things as did with a low NE (Jarvis and Petty, 1996). "High NE individuals are expected to engage in more evaluatively polarized thinking responding compared people who have low NE" (Fennis and Bakker, 2001). In the research of Fennis and Bakker this means for example that high NE persons may respond with extreme feelings of annoyance in reaction to an over-dramatized commercial (Fennis and Bakker, 2001). So, people with a high and low NE were exposed to a commercial. This resulted in that people with a high need to evaluate (NE) were: (1) more irritated after exposure to disliked or many ads and, consequently, (2) more negatively affected in their evaluations of the neutral ad and brand.

Irritation was measured by using a 7 items 5 category Likert scale. Participants were asked to indicate to what extent the commercials were annoying, irritating, boring, and bothersome. In addition, they were asked to what extent the commercials constituted a pleasant break from the program, distracted unpleasantly from the program, and formed a troublesome, irritating interference with the program (1=totally disagree, 5=totally agree). The reliability analysis showed a Cronbach's alpha of 0.87, which is very good. The irritation scale used by Fennis and Bakker is reliable.

Appendix 3F Irritation in working environment

In these studies the explained variances by the two factors (CI, EI) ranges from 16% to 58%. The reliability measured by Cronbach's alpha ranged from 0.84 to 0.93 which is quite good. The retest reliability is r =.69 after 6 months, r =.61 after 2.5 year and r = .57 after 3.5 year which is acceptable. Convergent correlation is measured by the coherence of the irritation-scale with other job-stressors. The irritation-scale is not compared with another scale also measuring irritation. So, if there a matter of good convergence validity is questionable. Discriminant validity is measured by the correlation of CI and EI with job-stressors. However the factors CI and EI correlate with each other r = .61 (Mohr et al., 2005). This high correlation implies that some items correlate high on both factors (like item 4, 5 and 8 see Appendix 3G). So, it can be concluded that discriminant validity is low. Their irritation scale can be used for evaluating interventions in the field of occupational health, for research on stress at work, and for individual counseling (Mohr et al., 2005). This scale is focused on measuring irritation in a work-environment, four of the eight items are related to your job (item 1,2, 4 and 8). Also the convergent and discriminant validity is questionable.

Appendix 3G Irritation-scale

Irritation-scale (in German)

Item	F1*	F2*
1. Es fällt mir schwer, nach der Arbeit abzuschalten. (KI)	.43	*.90*
2. Ich muss auch zu Hause an Schwierigkeiten bei der Arbeit denken. (KI)	.46	*.92*
3. Wenn andere mich ansprechen, kommt es vor, dass ich mürrisch reagiere. (EI)	*.79*	.46
4. Selbst im Urlaub muss ich manchmal an Probleme bei der Arbeit denken. (KI)	.50	*.85*
5. Ich fühle mich ab und zu wie jemand, den man als Nervenbündel bezeichnet. (EI)	*.79*	.52
6. Ich bin schnell verärgert. (EI)	*.87*	.38
7. Ich reagiere gereizt, obwohl ich es gar nicht will. (EI)	*.90*	.38
8. Wenn ich müde von der Arbeit nach Hause komme, bin ich ziemlich nervös. (EI)	*.76*	.51

KI = Kognitive Irration (Cognitive Irritation)
EI = Emotionale Irritation (Emotional Irritation)
* Factorload (explorative factor analysis, Promax Rotation)

Source: Mohr, Rigotti and Müller (2005)

Appendix 4A Description of symptoms

Symptoms	Description
Irritative symptoms	irritation of the nose, itching nose, dry nose, running nose, smarting nose, tiredness of the eyes, itching eyes, smarting eyes, irritation of the eyes, dry eyes watering eyes redness of the eyes, blurred sight, irritation of the throat and irritation of the skin
Olfactory symptoms	sensation of bad air, nasty smell, sensation of unpleasant smell, stink
Respiratory symptoms	pain or pressure over the chest coughing spells, shortness of breath

Source: Ihrig et al., 2006

Appendix 6A PANAS-X Dutch Version

Translation PANAS-X into Dutch

English	Dutch	English	Dutch
cheerful	vrolijk	active	actief
disgusted	walgen van	guilty	schuldbewust
attentive	oplettend	joyful	opgewekt
bashful	beduusd	nervous	nerveus
sluggish	lui	lonely	eenzaam
daring	moedig	sleepy	slaperig
surprised	verbaasd	excited	uitgelaten
strong	sterk	hostile	vijandig
scornful	minachtend	proud	trots
relaxed	ontspannen	jittery	zenuwachtig
irritable	prikkelbaar	lively	levendig
delighted	opgetogen	ashamed	beschaamd
inspired	geïnspireerd	at ease	op je gemak
fearless	onbevreesd	scared	angstig
disgusted with self	walgend van jezelf	drowsy	suf
sad	verdrietig	angry at self	boos op jezelf
calm	kalm	enthusiastic	enthousiast
afraid	bang	downhearted	terneergeslagen
tired	moe	sheepish	bedeesd
amazed	verwonderd	distressed	van streek
shaky	rusteloos	blameworthy	schuldig
happy	gelukkig	determined	aandachtig
timid	timide	frightened	bevreesd
alone	alleen	astonished	versteld staan
alert	alert	interested	geïnteresseerd
upset	overstuur	loathing	afkeer hebben van
angry	boos	confident	zelfverzekerd
bold	dapper	energetic	energiek
blue	zwaarmoedig	concentrating	geconcentreerd
shy	verlegen	dissatisfied with self	ontevreden met jezelf

Deze schaal bestaat uit een aantal woorden en korte zinnen die emoties en gevoelens beschrijven. Bekijk iedere beschrijving en geef een geschikt antwoord in de lege ruimte voor het woord. Geef hierbij aan in **welke mate jij je in het algemeen zo voelt**. Gebruik de volgende schaal om je antwoorden te beschrijven.

1: een heel klein beetje/helemaal niet
2: een beetje
3: gematigd
4: behoorlijk
5: extreem

1 ___	vrolijk	**31** ___	actief
2 ___	walgen van	**32** ___	schuldbewust
3 ___	oplettend	**33** ___	opgewekt
4 ___	beduusd	**34** ___	nerveus
5 ___	lui	**35** ___	eenzaam
6 ___	moedig	**36** ___	slaperig
7 ___	verbaasd	**37** ___	uitgelaten
8 ___	sterk	**38** ___	vijandig
9 ___	minachtend	**39** ___	trots
10 ___	ontspannen	**40** ___	zenuwachtig
11 ___	prikkelbaar	**41** ___	levendig
12 ___	opgetogen	**42** ___	beschaamd
13 ___	geïnspireerd	**43** ___	op je gemak
14 ___	onbevreesd	**44** ___	angstig
15 ___	walgend van jezelf	**45** ___	suf
16 ___	verdrietig	**46** ___	boos op jezelf
17 ___	kalm	**47** ___	enthousiast
18 ___	bang	**48** ___	terneergeslagen
19 ___	moe	**49** ___	bedeesd
20 ___	verwonderd	**50** ___	van streek
21 ___	rusteloos	**51** ___	schuldig
22 ___	gelukkig	**52** ___	aandachtig
23 ___	timide	**53** ___	bevreesd
24 ___	alleen	**54** ___	versteld staan
25 ___	alert	**55** ___	geïnteresseerd
26 ___	overstuur	**56** ___	afkeer hebben van
27 ___	boos	**57** ___	zelfverzekerd
28 ___	dapper	**58** ___	energiek
29 ___	zwaarmoedig	**59** ___	geconcentreerd
30 ___	verlegen	**60** ___	ontevreden met jezelf

Appendix 6B Use of PANAS-X

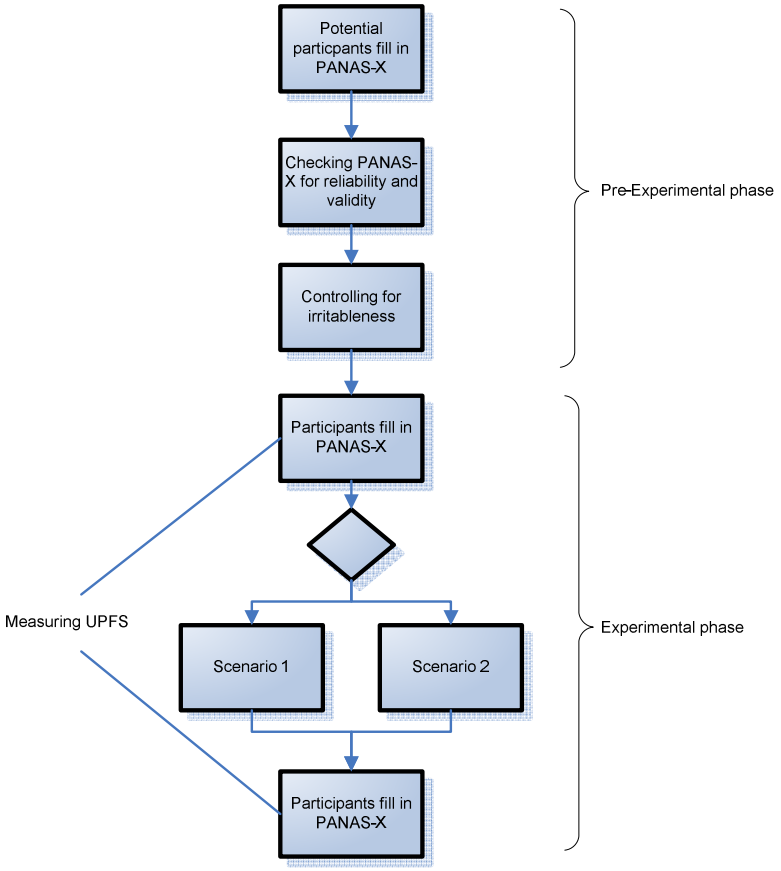


Figure 6B: Use of PANAS-X in this experiment

A short explanation of this figure will be given. But first a comment will made with regard to the Pre-Experimental phase. The activities related to the pre-experimental phase of this research project are combined with the pre-experimental phase of Gielen's research (Gielen, 2007). As said, the goal of Gielen's research project is to investigate to what extent function importance influences UPFS. For reaching this goal, Gielen also had to apply some selection criteria on the potential participants. For determining what a high and low importance function is and for selecting students a Function Importance Questionnaire was developed.

Returning to the first phase: "potential students fill in PANAS-X" of figure 6B. Potential participants were approached with the question if they want to participate in an experiment. They were asked to fill in the questionnaire which contained the PANAS-X schedule, the Function Importance Questionnaire, some general questions and the question whether he/she (maybe) wants to participate in an experiment. The general questions concern age, gender, study-discipline, cost, age and functions of his/her television. The general questions were especially useful for the master thesis project of Gielen. For this master thesis project (of Bergmans), these general questions were used for checking possible differences in irritableness and UPFS between groups for example gender or study discipline. From the 130 participants who filled in the (PANAS-X) questionnaire, 123 questionnaires are useable. Reasons for excluding 7 questionnaires were: a not completed questionnaire (missing more than 5 items); the questionnaire was not filled in with a 5 point Likert scale, instead of using a scale ranging from 0-5 a scale was used ranging from 0-10; or the filled in questionnaire was not readable. Also 77 similar PANAS-X questionnaires from another master

thesis project of Yvonne Kleuskens (2007) were used for selecting participants. These questionnaires are added for creating a larger potential group of participants. These participants were also Dutch students of the Eindhoven University of Technology and part of the same homogeneous group. In total 200 students successfully completed the PANAS-X questionnaire.

Appendix 6C Data examination

In total 9 values of the 200*60 = 12.000 values were missing (< 1%). It is remarkable that 8 of these missing values are missed in the affective state Shyness. Four times a value was missed at Sheepish, three times a value was missed at Timid and one time a value was missed at Bashful. This could be an indication that the affective state Shyness is not translated very well. Apparently Dutch students do not understand the Dutch translation of Sheepish, Timid and Bashful. These missing values are replaced. When the missing data is under 10% any imputation method can be applied (Hair et al., 2005). Mean Value Regression method is used as imputation method. This method predicts the missing values of a variable based on its relationship with other variables in the data set (Hair et al., 2005). So, the missing values are "replaced" with the predicted values and a complete dataset of 200 cases will be used for further analysis.

Appendix 6D Comparison with Maastricht Aging Study

The analysis of the Failure Severity Study is done in exactly the same as in the Maastricht Aging Study. Only the 20 items which are relevant for the general dimensions positive affect and negative affect are kept in the database of the Failure Severity Study for this analysis. Two factors are extracted out of this data in order to get the Negative Affect "factor" and the Positive Affect "factor". The analysis method and rotation method used is the same as used in the Maastricht Aging Study: principal component analysis and varimax rotation. The appropriateness of these analysis and rotation methods will be discussed further on. Now the focus is on comparing the results of the Failure Severity Study and Maastricht Aging Study. This comparison can demonstrate that the 20 items of PANAS (more specifically 20 items of the 60 items consisting PANAS-X) are used in a reliable and valid way. The results of this factor analyses for both studies are presented in table 6D1, presented on the next page.

Both analyses produced an underlying factor structure which is inline with the factor structure described by Watson et al. (1988), namely a positive affect factor and a negative affect factor that together encompassed all 20 PANAS items. A visual inspection of table 5.3.2.1 shows that there is a large difference (compared with other items) in the mean score of the items guilty, proud and excited. It has to be noted that the mean age of the participants of the Maastricht Aging is 62.67 years where in the Failure Severity Study the mean age is around 22[4]. Apparently these younger people (students) feel more guilt than older people. Students also are more proud and excited than older people. The two factors explain in both studies nearly the same percentage of variance (around 38%). The internal consistencies of the factors of both studies are around the 0.80 (Cronbach's alpha). However it has to be noted that Cronbach's alpha of the PA factor is lower in the Failure Severity Study (0.75) than the Maastricht Aging Study (0.84). The reliability of PA is lower in the Failure Severity Study, but is still above the minimum recommended level of 0.7. Besides these differences the results are nearly the same for both studies. These results confirm that the 20 items forming PANAS are reliable and can be used for further analysis of irritableness and UPFS.

[4] Only the persons who were willing to participate in the experiment filled in their year of birth (N=112).

Table 6D1: Principal components analysis of PANAS items with varimax rotation*

Item	Failure Severity Study			Maastricht Aging Study		
	mean (SD)	Factor 1(**NA**)	Factor 2 (**PA**)	mean (SD)	Factor 1 (**NA**)	Factor 2 (**PA**)
afraid	1.50 (.716)	0.72		1.73 (1.07)	0.62	
scared	1.51 (.736)	0.71		1.63 (0.95)	0.67	
distressed	1.43 (.661)	0.71		1.76 (1.14)	0.73	
jittery	2.03 (.847)	0.70		2.36 (1.30)	0.68	
nervous	2.04 (.841)	0.69		2.44 (1.23)	0.75	
upset	1.63 (.904)	0.65		1.86 (1.11)	0.78	
ashamed	1.80 (.835)	0.64		1.47 (0.82)	0.51	
irritable	2.25 (.975)	0.60		2.27 (1.08)	0.62	
hostile	1.60 (.897)	0.59		1.34 (0.74)	0.36	
guilty	2.79 (1.120)	0.40		1.51 (0.90)	0.55	
enthusiastic	3.57 (.767)		0.69	3.24 (1.05)		0.71
active	3.54 (.929)		0.67	3.73 (1.08)		0.66
alert	3.30 (.833)		0.65	3.32 (1.11)		0.54
determined	3.46 (.861)		0.61	3.40 (1.08)		0.61
interested	3.68 (.728)		0.60	3.90 (0.91)		0.55
inspired	3.35 (.811)		0.60	3.06 (1.10)		0.70
proud	3.22 (.902)		0.48	2.54 (1.22)		0.56
excited	2.92 (.893)		0.44	2.20 (1.19)		0.38
attentive	3.42 (.810)		0.43	3.41 (1.01)		0.55
strong	3.22 (.902)		0.39	3.11 (1.12)		0.64
Eigenvalues		4.37	3.12		4.12	3.58
% of variance explained		21.87	15.62		20.58	17.89
% of cumulative variance explained		21.87	37.49		20.58	38.46
Cronbach's alpha		0.83	0.75		0.80	0.84

* loadings < |0.30| are not shown

Appendix 6E Factor analysis

In addition to the two higher order scales there are also 11 specific affects within PANAS-X (not within PANAS). These 11 specific affects can be categorized in three categories: Basic Negative Emotion, Basic Positive Emotion and Other Affective States, figure 6E1. The scales Fear, Hostility, Guilt and Sadness form together the Basic Negative Emotion Scale. These scaled are also related to the General Dimension scale Negative Affect. Consequently, the Basic Negative Emotion Scale is also related to the General Dimension scale Negative Affect. The same reasoning can be done for the Basic Positive Emotion Scales. Until now it is not clear whether the states Shyness, Fatigue Serenity or Surprise are related to Negative or Positive Affect (or a combination of Negative and Positive Affect). All these 11 specific affect states are composed of a number of items, see table 6E1 (see also appendix 3C). For example, the specific affect state Fear consists of the following items: afraid, scared, frightened, nervous, jittery and shaky. All together PANAS-X is made up of 60 items. These 60 items together form the 11 specific affects and the two general Positive and Negative Affects. The score on the General Dimension Scale Negative Affect (or Positive Affect) is calculated by summing the scores of the 10 accompanying items. The 11 specific affect states are also calculated by adding the scores of the accompanying items. For Basic Negative Emotion, the scores of the specific affect scales, fear, hostility, guilt and sadness is summed and consequently divided by 4. For Basic Positive Emotion, the score on the specific affect scales, joviality,

self-assurance and attentiveness is added and consequently divided by 3, see table 6E1. PANAS-X can be described as "a hierarchical taxonomic scheme in which two broad, higher order dimensions are each composed of several correlated, yet ultimately distinguishable states" (Watson and Clark, 1999). With factor analysis it will be assessed whether the proposed structure also exist within the Dutch version of PANAS-X. With this factor analysis the two higher order dimensions, Positive and Negative Affect, are assessed. Also the 11 specific states will be examined. First however, the assumptions underlying factor analysis will be checked.

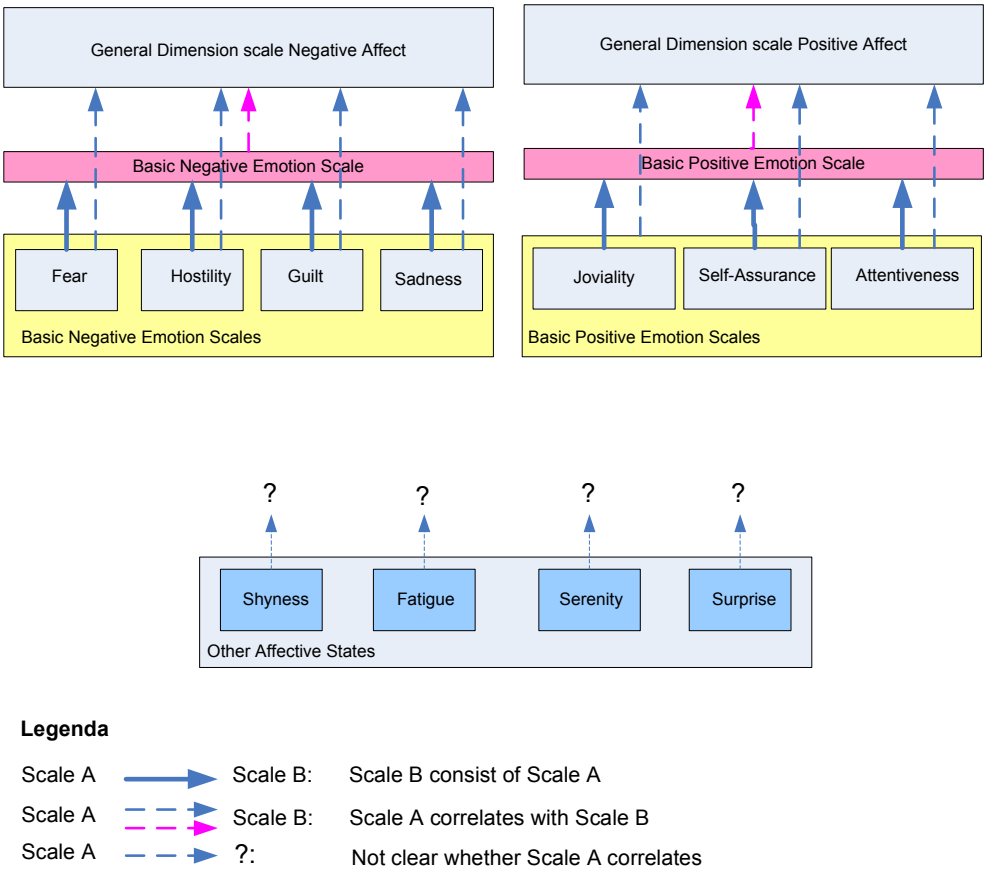


Figure 6E1: Hierarchical structure of PANAS-X

Table 6E1: Item composition of the PANAS-X scales

Scales	**General Affects**	**Calculation of score**
General Dimensions Scale	Negative Affect	afraid + scared + nervous+ jittery + irritable + hostile + guilty + ashamed + upset + distressed
General Dimensions Scale	Positive Affect	active + alert + attentive + determined + enthusiastic + excited + inspired + interested + proud + strong
	Specific Affects	
Basic Negative Emotion Scale		(Fear + Hostility + Guilt + Sadness) / 4
	Fear	afraid + scared + frightened + nervous + jittery + shaky
	Hostility	angry + hostile + irritable + scornful + disgusted + loathing
	Guilt	guilty + ashamed + blameworthy + angry at self + disgusted with self + dissatisfied with self
	Sadness	sad + blue + downhearted + alone + lonely
Basic Positive Emotion Scale		(Joviality + Self-Assurance + Attentiveness) / 3
	Joviality	happy + joyful + delighted + cheerful + excited+ enthusiastic + lively + energetic
	Self-Assurance	proud + strong + confident + bold + daring + fearless
	Attentiveness	alert + attentive + concentrating + determined
Other Affective States		
	Shyness	shy + bashful + sleepish + timid
	Fatigue	sleepy + tired + sluggish + drowsy
	Serenity	calm + relaxed + at ease
	Surprise	amazed + surprised + astonished

Appendix 6F Assumptions underlying factor analysis

The assumptions underlying factor analysis will be analyzed. The critical assumptions underlying factor analysis are more conceptual than statistical (Hair et al., 2005). A basic conceptual assumption underlying factor analysis is that some underlying structure exists in the set of selected variables (Hair et al., 2005). The existence of this underlying structure is widely proven by Watson, Clark and Tellegen. A statistical issue is to ensure that the variables are sufficiently intercorrelated to produce representative factors (Hair et al., 2005). When all of the correlations are low, or equal then it is questionable whether factor analysis is applicable. For testing the applicability of factor analysis two methods are widely used: Bartlett test of sphericity and the Measure of Sampling Adequacy (MSA). A Bartlett test of sphericity tests whether there is sufficient correlation among the variables. The measure of sample adequacy (MSA) quantifies the degree of intercorrelations among the variables and the appropriateness of factor analysis. This index ranges from 0 to 1, reaching 1 when each variable is perfectly predicted without error by other variables. The MSA values must exceed .5 for both the overall test and each individual variable (Hair et al., 2005), for showing a sufficient degree of intercorrelations among the variables. The overall MSA test shows a value of 0.812 which is meritorious (Hair et al., 2005) and largely exceeds 0.5, see table 6F1. Also the Bartlett test of sphericity is highly significant, see table 6F1. From this test it can be concluded that there are significant correlations among the variables. The MSA values of the individual items do all pass the criteria of 0.5, see table 6F2. Each item seems to fit within the structure of the other items. Overall there is sufficient degree of intercorrelations between variables which indicates that a factor analysis may be useful with this data.

Table 6F1: MSA and Bartlett's Test

Measure of Sampling Adequacy		0.812
Bartlett's Test of Sphericity	Chi-Square	5480.413
	df	1770
	Sig.	0.00

Table 6F2: MSA of individual items

	MSA		MSA		MSA		MSA
Bang	0.840	Enthousiast	0.825	Boos op jezelf	0.879	Geïnteresseerd	0.718
Angtig	0.853	Levendig	0.848	Walgend van jezelf	0.863	Verlegen	0.819
Bevreesd	0.843	Energiek	0.840	Ontevreden met jezelf	0.822	Beduusd	0.797
Nerveus	0.882	Trots	0.729	Verdrietig	0.836	Bedeesd	0.790
Zenuwachtig	0.900	Sterk	0.669	Zwaarmoedig	0.883	Timide	0.758
Rusteloos	0.790	Zelfverzekerd	0.677	Teneergeslagen	0.889	Slaperig	0.801
Boos	0.909	Dapper	0.717	Alleen	0.802	Moe	0.797
Vijandig	0.811	Moedig	0.756	Eenzaam	0.816	Lui	0.796
Prikkelbaar	0.840	Onbevreesd	0.691	Overstuur	0.914	Suf	0.833
Minachtend	0.729	Alert	0.684	Van streek	0.881	Kalm	0.523
Walgen van	0.790	Oplettend	0.589	Gelukkig	0.777	Ontspannen	0.725
Afkeer hebben van	0.817	Geconcentreerd	0.768	Opgewekt	0.719	Op je gemak	0.759
Schuldbewust	0.737	Aandachtig	0.591	Opgetogen	0.770	Verwonderd	0.778
Beschaamd	0.897	Actief	0.862	Vrolijk	0.776	Verbaasd	0.834
Schuldig	0.830	Geïnspireerd	0.733	Uitgelaten	0.808	Versteld staan	0.864

Appendix 6G Performing factor analysis

PANAS-X consists out of two broad higher dimensions: Negative Affect and Positive Affect. As said before: PANAS-X can be described as "a hierarchical taxonomic scheme in which two broad, higher order dimensions are each composed of several correlated, yet ultimately distinguishable states" (Watson and Clark, 1999). With this factor analysis these two higher order dimensions are examined. This analysis is done for assessing the general structure of PANAS-X. The objective is to extract two factors, Negative Affect and Positive Affect. The objective is to show that these extracted factors measure Positive Affect and Negative Affect as proposed by Watson and his colleagues (Watson and Clark, 1999). This means that all items which are related to the Negative Affect should load on one factor. These are the items of the General Dimension Scale Negative Affect and all items of the Basic Negative Emotion Scale. All items which are related to Positive Affect should load on the other factor. For the items related to other affective state no clear pattern is expected. These items can load on any factor.

There are two methods for doing factor analysis: common factor analysis and component analysis. Component analysis is especially suitable for summarizing most of the original information in a minimum number of factors for prediction purposes (Hair et al., 2005). Common factor analysis is used primarily to identify underlying factors or dimensions that reflect what the variables share in common (Hair et al., 2005). Common factor analysis has two problems. With common factor analysis it is possible that several different factor scores[5] can be calculated from a single factor model. No single unique solution is found, as in component analysis (Hair et al., 2005). The second issue is that sometimes the communalities are not estimable or may be invalid, requiring the deletion of the variable from the analysis (Hair et al., 2005). Communality is the total amount of variance an original variable shares with all other variables included. There is some debate over which factor model is more appropriate, empirical research shows that similar results are obtained in many instances when using component or common factor analysis (Hair et al., 2005). Especially if the numbers of variables exceeds 30 (Hair et al., 2005).

Two factors will be extracted out of the 60-items of PANAS-X. The objective is to examine the underlying structure for two broad dimensions namely Positive Affect and Negative Affect. This unrotated factor solution will be rotated for improving the interpretation. The goal of rotation is to derive theoretically meaningful factors and, if possible, the simplest factor structure (Hair et al., 2005). Figure 6G1 shows as factor rotation. The reference axes of the factors are turned about the origin in such a way that a simpler, theoretically more meaningful factor pattern is reached (Hair et al., 2005).

[5] A factor score is a composite measure created for each observation on each factor extracted in the factor analysis. The factor weights are used in conjunction with the original variable value to calculate each observation's score. The factor score can be used to represent the factor(s) in subsequent analysis. Factor scores are standardized to have a mean of 0 and a standard deviation of 1. Source Hair et al. (2005).

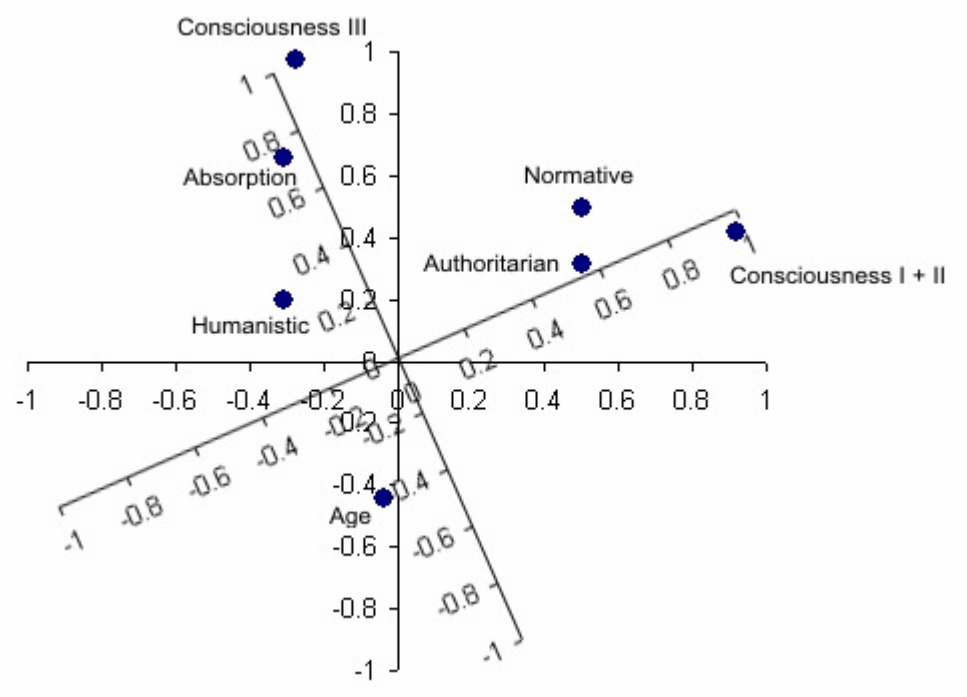


Figure 6G1: Example of factor rotation , source Krus D.J. & Tellegen A. (1975)

There are two rotation methods orthogonal and oblique. In orthogonal factor rotation the axes will be maintained at 90 degrees (figure 6G1). Within oblique rotation the axes do not have to be retained at 90 degrees. There are no specific rules for preferring one factor method over another (Hair et al., 2005). Orthogonal rotation methods are most widely used rotation methods by researchers. The three most known orthogonal rotation methods are Quartimax, Varimax and Equimax. The objective of all rotation methods is to simplify the rows and columns of the factor matrix to increase the interpretation. In a factor matrix, columns represent factors with each row corresponding to a variable's loading across the factors. By simplifying the rows, as many values in each row are made as close to zero as possible. A variable's loading on a single factor is maximized. By simplifying the columns, as many variables in each column are made as close to zero as possible. Making the number of high loadings as few as possible (Hair et al., 2005). There are three widely known orthogonal approaches: Quartimax, Varimax and Equimax. The goal of Quartimax is to simplify the rows of a factor matrix. Quartimax has not been proven very successful in producing simpler structures (Hair et al., 2005). Varimax concentrates on simplifying the columns of the factor matrix. The Varimax method has proved to be successful in producing a clearer separation of factors (Hair et al., 2005) than Quartimax. Varimax tries to derive high loadings -1 or +1 in each column of the matrix, thus indicating a clear positive or negative association between variable and factor; or 0, indicating a lack of association (Hair et al., 2005). Equimax tries to simplify both rows and columns of a factor matrix. Equimax has not gained widespread acceptance and is used infrequently. The most popular orthogonal rotation method in literature is Varimax (Conway and Huffcutt, 2003).Therefore Varimax rotation method will be used.

Appendix 6H Principal components analysis of PANAS-X items with varimax rotation 1

			Component					Component	
item	Mean	SD	1	2	item	Mean	SD	1	2
Bang	1.50	0.716	0.623		Enthousiast	3.57	0.767		0.678
Angstig	1.51	0.736	0.650		Levendig	3.59	0.778		0.654
Bevreesd	1.64	0.840	0.660		Energiek	3.56	0.819		0.747
Nerveus	2.04	0.841	0.631		Trots	3.22	0.856		0.416
Zenuwachtig	2.03	0.847	0.647		Sterk	3.22	0.902		0.413
Rusteloos	1.81	1.016	0.480		Zelfverzekerd	3.43	0.818		0.456
Boos	1.69	0.841	0.613		Dapper	2.85	0.950		0.490
Vijandig	1.60	0.897	0.549		Moedig	3.12	0.903		0.541
Prikkelbaar	2.25	0.975	0.577		Onbevreesd	2.95	0.955		0.421
Minachtend	1.73	0.966	0.376		Alert	3.30	0.833		0.533
Walgen van	1.50	0.783	0.495		Oplettend	3.42	0.810		
Afkeer hebben van	1.84	0.934	0.545		Geconcentreerd	3.24	0.810		0.489
Schuldbewust	2.79	1.120	0.334		Aandachtig	3.46	0.861		0.374
Beschaamd	1.80	0.835	0.599		Actief	3.54	0.929		0.631
Schuldig	1.63	0.798	0.613		Geïnspireerd	3.25	0.811		0.506
Boos op jezelf	1.59	0.870	0.615		Geïnteresseerd	3.68	0.728		0.408
Walgend van jezelf	1.32	0.656	0.626		Verlegen	2.36	0.941	0.352	
Ontevreden met jezelf	1.64	0.846	0.526		Beduusd	1.86	0.807	0.376	
Verdrietig	1.58	0.810	0.601		Bedeesd	1.87	0.814	0.412	
Zwaarmoedig	1.74	0.816	0.566		Timide	2.28	1.042	0.356	
Teneergeslagen	1.52	0.694	0.593		Slaperig	2.48	1.080	0.505	
Alleen	1.78	0.869	0.560		Moe	2.67	1.080	0.506	
Eenzaam	1.59	0.816	0.590	-0.343	Lui	2.53	1.093		
Overstuur	1.63	0.904	0.599		Suf	1.93	0.961	0.474	
Van streek	1.43	0.661	0.715		Kalm	3.47	0.924		
Gelukkig	3.92	0.675		0.510	Ontspannen	3.51	0.845		
Opgewekt	3.68	0.678		0.575	Op je gemak	3.73	0.692		0.345
Opgetogen	3.36	0.789		0.451	Verwonderd	2.25	0.917	0.467	0.322
Vrolijk	3.82	0.658		0.533	Verbaasd	2.15	0.893	0.417	
Uitgelaten	2.92	0.893		0.437	Versteld staan	2.20	0.939	0.529	

* loadings < |0.30| are not shown

Appendix 6I Overall structure PANAS-(X)

The overall structure of the two broad higher dimensions is confirmed. In this analysis, it also became visible that most items load on the high pole ends, see figure 6I1. This is in line with Watson et al. who claim that the high pole ends are much more important than the low ends (Watson et al., 1999). All important instruments that assess mood contain many more high-activation terms (located at high pole ends) than low activation terms (Watson et al., 1999). Also, the high pole ends (the activated ends) define the dimensions. The low poles of these dimensions present the absence of a particular activation (Watson et al., 1999).

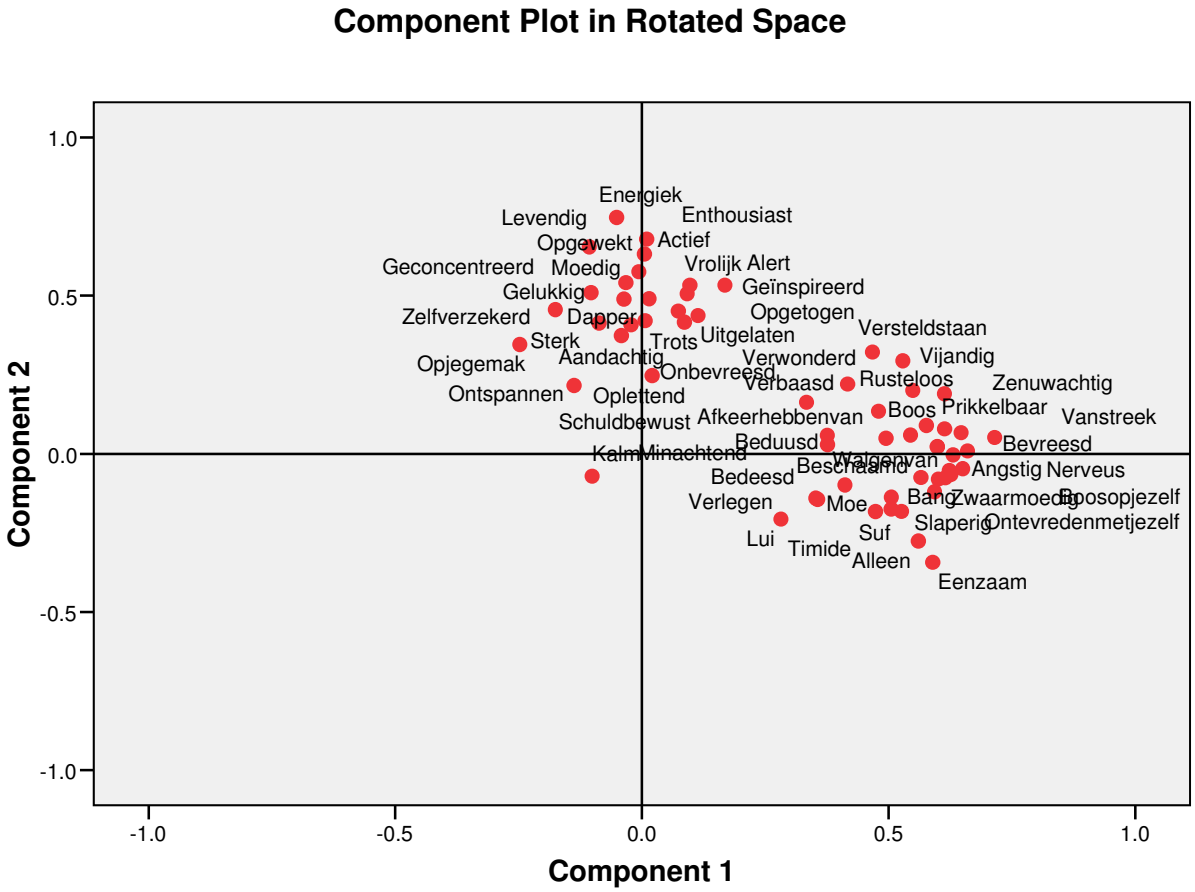


Figure 6I1: Component plot factor 1 (component 1) and factor 2 (component 2)

Appendix 6J Principal components analysis of PANAS-X items with varimax rotation 2

		Component					
English	Dutch	1	2	3	4	5	6
distressed	van streek	0.705					
disgusted with self	walgend van jezelf	0.686					
angry at self	boos op jezelf	0.650					
frightened	bevreesd	0.648					
blameworthy	schuldig	0.626					
ashamed	beschaamd	0.604					
jittery	zenuwachtig	0.591					
angry	boos	0.585					
downhearted	terneergeslagen	0.585					
upset	overstuur	0.576					
sad	verdrietig	0.571					
lonely	eenzaam	0.567	-0.455				
nervous	nerveus	0.566					
scared	angstig	0.563					
dissatisfied with self	ontevreden met jezelf	0.561					
irritable	prikkelbaar	0.557					
astonished	versteld staan	0.554					
loathing	afkeer hebben van	0.549					
blue	zwaarmoedig	0.543					
disgusted	walgen van	0.540					
afraid	bang	0.537					
alone	alleen	0.527	-0.402				
hostile	vijandig	0.524					
shaky	rusteloos	0.519					
surprised	verbaasd	0.459					
amazed	verwonderd	0.458					
sheepish	bedeesd	0.401					
guilty	schuldbewust						
scornful	minachtend						
bashful	beduusd						
energetic	energiek		0.685				
joyful	opgewekt		0.670				
cheerful	vrolijk		0.631				
lively	levendig		0.620				
enthusiastic	enthousiast		0.611	0.441			
happy	gelukkig		0.501				
excited	uitgelaten		0.440				
active	actief		0.405				
delighted	opgetogen						
at ease	op je gemak						
determined	aandachtig			0.710			
attentive	oplettend			0.609			
alert	alert			0.560			
interested	geïnteresseerd			0.560			
concentrating	geconcentreerd			0.520			
inspired	geïnspireerd			0.427			
proud	trots						
daring	moedig				0.722		
bold	dapper				0.645		
strong	sterk				0.625		
fearless	onbevreesd				0.624		
confident	zelfverzekerd						
tired	moe					0.762	
sleepy	slaperig					0.726	
sluggish	lui					0.638	
drowsy	suf					0.598	
calm	kalm						0.538
timid	timide						0.533

shy	verlegen							0.520
relaxed	ontspannen							0.410

* loadings < |0.40| are not shown

This discussion starts with the results of the second factor. The first factor will be addressed later on. The items which positively load on the second factor are energetic, joyful, cheerful, lively, enthusiastic, happy, excited and active. These items are all related to positive affect. With the exception of the item active all these items are part of the basic emotion scale Joviality. Therefore the second factor could be seen as the basic emotion Joviality. The items lonely and alone load negatively on this second factor. This negative loading could be explained by the fact these two items are related to negative affect, the "opposite" of positive affect where Joviality belongs to. The third factor contains the items enthusiastic, determined, attentive, alert, interested, concentrating and inspired. The items alert, attentive, concentrating and determined belong to the basic positive emotion Attentiveness. The other three items enthusiastic, interested and inspired belong to the general dimension scale positive affect. So, the third factor could be seen as the basic emotion Attentiveness. The fourth factor contains the items daring, bold, strong and fearless. These four items are part of the basic emotion Self-Assurance. Therefore the fourth factor can be seen as the basic emotion Self-Assurance. The fifth factor consists out of the items tired, sleepy, sluggish and drowsy. These items are exactly the items of the "other" affective state Fatigue. For that reason the fifth factor can be seen as Fatigue. The sixth factor contains the items calm, timid, shy and relaxed. These items are part of the other affective state Shyness and Serenity. This sixth factor shows no clear pattern. Returning to the first factor, it is visible that nearly all items (except, scornful and guilty) of the basic negative emotion scales Fear, Hostility, Guilt and Sadness load on this factor. So, this factor contains all items related to the basic negative emotion scales. An attempt was made for specifying this factor further. This was done by selecting all the items of factor 1 and the items scornful and guilty and deleting the items of the other affective states which loaded on factor 1. Several solutions were examined. The best interpretable solution is a five factor solution but this solution does not contain the structure as proposed by Watson and Clark (Watson and Clark, 1999) (see also figure 6J1). However in figure 6J1 it is visible that only the items related to Surprise load on only one factor. So, the existence of the Surprise scale could be extracted out of the dataset. It is not possible to extract the specific basic negative emotions. This finding is in line with the results presented by Bagozzi (1993). Bagozzi reanalyzed the data in Watson and Clark (1991, 1992). Bagozzi found that when someone wants to examine the overall negative affect, the composite of the basic negative affect scales may be useful to employ. This composite is the Basic Negative Emotion Scale which consists out of the scales Fear, Hostility, Guilt and Sadness. However, if one wants to examine the unique contributions of these specific basic negative affect scales (Fear, Hostility, Guilt, and Sadness), the individual basic scales are useful only when well-defined conditions can be identified that lead to differential emotional responses (Bagozzi, 1993). This conclusion is based on the mixed findings for discriminant validity of the basic negative emotion scales.

Rotated Component Matrix(a)

	Component				
	1	2	3	4	5
Bang				.704	
Angtig				.769	
Bevreesd			.460		
Nerveus			.415		
Zenuwachtig					
Rusteloos		.429			
Boos		.518		.403	
Vijandig		.693			
Prikkelbaar		.545			
Minachtend		.716			
Walgen van		.556			
Afkeer hebben van		.490			
Schuldbewust			.655		
Beschaamd			.637		
Schuldig			.621		
Boos op jezelf	.447		.439		
Walgend van jezelf	.638				
Ontevreden met jezelf	.525		.521		
Verdrietig	.636				
Zwaarmoedig	.452				
Teneergeslagen	.445				
Alleen	.790				
Eenzaam	.798				
Verwonderd					.812
Verbaasd					.701
Versteld staan					.691

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 7 iterations.

Figure 6J1: Rotated Component Matrix Basic Negative Emotions

Appendix 6K Validity of PANAS-X

Besides reliability also the validity of PANAS-X has to be tested. Validity is the extent to which a scale or set of measures accurately represents the concept of interest (Hair et al., 2005). There are several forms of validity. Two widely used measures of validity are convergent validity and discriminant validity. Convergent validity assesses the degree to which two measures of the same concept are correlated (Hair et al., 2005). Discriminant validity is the degree to which a scale is sufficiently different from other related scales (Hair et al., 2005). For assessing convergent and discriminant validity the correlation matrix is given below, table 6K1. For showing convergent validity the correlations have to be high between related scales. For discriminant validity, the correlations have to be low between unrelated scales. The matrix of 6K1 will be discussed below.

Table 6K1 Correlation matrix scales PANAS-X

	Fear	Hostility	Guilt	Sadness	Joviality	Self-Assurance	Attent-iveness	Shyness	Fatigue	Serenity	Surprise	General Pos Aff	General Neg Aff	Bas Pos Emo	Bas Neg Emo
Fear															
Hostility	0.614*														
Guilt	0.654*	0.524**													
Sadness	0.589*	0.502**	0.580**												
Joviality				0.229**											
Self-Assurance					0.414**										
Attentiveness					0.352**	0.281**									
Shyness	0.450**	0.260**	0.368**	0.510**											
Fatigue	0.481**	0.373**	0.313**	0.420**				0.293**							
Serenity	-0.187**	-0.262**	-0.212**	-0.220**	0.156*	0.168*									
Surprise	0.535**	0.413**	0.430**	0.302**	0.214**		0.154*	0.350**							
General Pos Aff					0.655**	0.492**	0.747**			0.160*	0.251**				
General Neg Aff	0.870**	0.741**	0.768**	0.634**				0.430**	0.450**	-0.237**	0.509**				
Bas Pos Emo				-0.205**	0.800**	0.820**	0.609**		-0.160*	0.204**	0.187**	0.793**			
Bas Neg Emo	0.867**	0.812**	0.836**	0.787**				0.471**	0.479**	-0.266**	0.513**		0.916**		

** correlation is significant at the 0.01 level (2-tailed)

* correlation is significant at the 0.05 level (2-tailed)

Positive and Negative Affect

The convergent and discriminate validity of the basic negative (Fear, Hostility, Guilt and Sadness) and basic positive emotion scales (Joviality, Self-Assurance and Attentiveness) will be examined. Table 6K2 will serve as a guide for exploring the correlations between these scales.

Table 6K2: Guide for correlations 1

	Fear	Hostility	Guilt	Sadness	Joviality	Self-Assurance	Attent-iveness
Fear							
Hostility	0.614*						
Guilt	0.654*	0.524**					
Sadness	0.589*	0.502**	0.580**				
Joviality				0.229**			
Self-Assurance					0.414**		
Attentiveness					0.352**	0.281**	
Shyness							
Fatigue							
Serenity							
Surprise							
General Pos Aff					0.655**	0.492**	0.747**
General Neg Aff	0.870**	0.741**	0.768**	0.634**			
Bas Pos Emo				-0.205**	0.800**	0.820**	0.609**
Bas Neg Emo	0.867**	0.812**	0.836**	0.787**			

** correlation is significant at the 0.01 level (2-tailed)
* correlation is significant at the 0.05 level (2-tailed)

First the correlations are examined in the pink block. The emotion scales Fear, Hostility, Guilt and Sadness are correlated very significantly. These three scales are also correlated very significantly and highly with the general negative affect and basic negative emotion scale (green block). These three scales (pink block) are not significantly correlated with the basic positive emotion scales Joviality, Self-Assurance and Attentiveness except for the sadness scale. The Sadness scale is significantly correlated with Joviality, but the size of the correlation is not that high (0.229) in comparison with the correlation of Sadness with Fear, Hostility and Guilt. Sadness is also significantly correlated with the basic positive emotion scale. It is important to note that the direction of this correlation is negative. The correlation between the basic positive emotion scales Joviality, Self-assurance and Attentiveness is significant (purple block), but the size is (much) lower than the correlation among the basic negative scales (Fear, Hostility, Guilt and Sadness, pink block). This is also a reason why factor analysis could extract three factors for the three basic positive emotion scales Joviality, Self-Assurance and Attentiveness, but could only extract one factor for the four negative emotion scales Fear, Guilt, Hostility and Sadness. Joviality, Self-assurance and Attentiveness (purple block) load high on general positive affect and basic positive emotion scale (yellow block). These results show that there is sufficient discriminant validity between the positive and negative emotion scales, because there is (nearly) no correlation between these emotion scales. The high correlations of these scales with their corresponding higher order factor show that there is sufficient convergent validity. The scales indeed measure positive or negative affect. The discriminant validity of the scales Fear, Hostility, Guilt and Sadness is questionable, because the correlation of these emotion scales is high (pink block). The scales Fear, Hostility, Guilt and Sadness have a high communality. It is questionable whether Fear, Hostility, Guilt and Sadness really measure something different or just stick together measuring one emotion. The correlation between Joviality, Self-assurance and Attentiveness (purple block) is not that big as in the case of Fear, Hostility, Guilt and Sadness (pink

block). This smaller correlation implies that Joviality, Self-assurance and Attentiveness (purple block) measure a different concept in such a way that discriminant validity is sufficient.

Other affective states

The convergent and discriminate validity of the other affective states (Shyness, Fatigue, Serenity and Surprise) will be examined. Table 6K3 will be used as a guide for exploring the relevant correlations with regard to the other affective states.

Table 6K3: Guide for correlations 2

	Shyness	Fatigue	Serenity	Surprise	
Fear	0.450**	0.481**	0.187**	0.535**	1
Hostility	0.260**	0.373**	0.262**	0.413**	
Guilt	0.368**	0.313**	0.212**	0.430**	
Sadness	0.510**	0.420**	0.220**	0.302**	
Joviality			0.156*	0.214**	2
Self-Assurance			0.168*		
Attentiveness				0.154*	
Shyness		0.293**		0.350**	3
Fatigue	0.293**				
Serenity					
Surprise	0.350**				
General Pos Aff			0.160*	0.251**	4
General Neg Aff	0.430**	0.450**	0.237**	0.509**	
Bas Pos Emo		0.160*	0.204**	0.187**	
Bas Neg Emo	0.471**	0.479**	0.266**	0.513**	

** correlation is significant at the 0.01 level (2-tailed)
* correlation is significant at the 0.05 level (2-tailed)

Shyness

Shyness correlates very significantly with the negative basic emotions Fear, Hostility, Guilt and Sadness (especially with Fear and Sadness, block 1, pink). Shyness correlates very significantly with general negative affect and the basic negative emotion scale (block 4, pink). Shyness correlates not only with the "negative affect scales" but also with Fatigue and Surprise (Block 3, pink). These correlations are not negligibly (0.293 and 0.350). So, the claim that the Shyness scale does possess sufficient discriminant validity can not be proven.

Fatigue

Fatigue correlates very significantly with Fear, Hostility Guilt and Sadness, especially with Fear (block 1, purple). Fatigue correlates very significantly with general negative affect and the basic negative emotion scale (block 4, purple). Fatigue correlates substantially positively with Shyness (block 3, purple) and negatively with the basic positive emotion scale (block 4, purple). Also in this case for Fatigue, the substantial correlation with Shyness is a reason to reject the idea that Fatigue- scale posses sufficient discriminant validity.

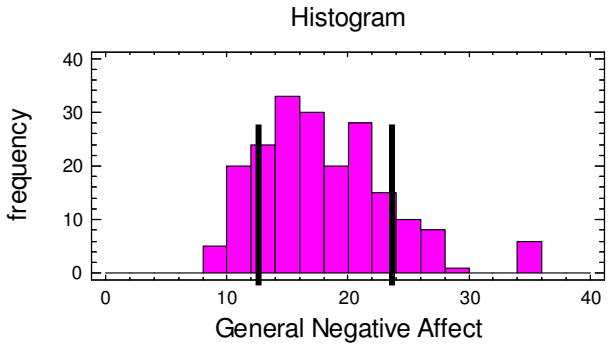
Serenity

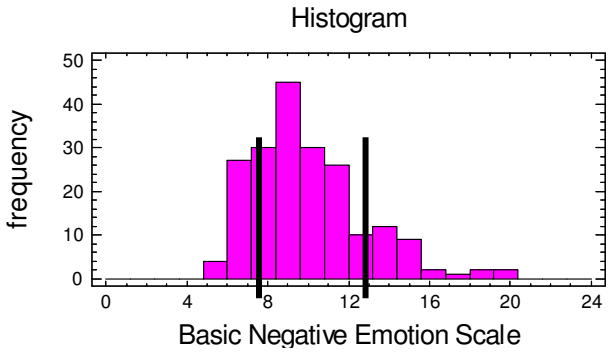
Serenity correlates negatively with Fear, Hostility Guilt and Sadness but not that high as the other affective states (Shyness, Fatigue and Surprise), (block 1). Serenity also correlates positively with Joviality and Self-Assurance (block 2, green). Serenity correlates positively with general positive affect and the basic positive emotion scale and negatively with general negative affect and the basic negative emotion (block 4, green). Of course, the Serenity scale is not valid because it is not even reliable.

Surprise

Surprise correlates positively with a lot of scales (Fear, Hostility, Guilt, Sadness, Joviality, Attentiveness and Shyness, general positive affect, general negative affect, general negative affect, basic positive emotion scale and basic negative emotion, (all blocks, yellow). So, it is very questionable if Surprise is sufficiently different from the other scales. It is not clear what Surprise measures, so discriminant validity is not sufficient.

Appendix 6L Histogram General Negative Affect and Basic Negative Emotion





Appendix 7A Experimental room

Error! Objects cannot be created from editing field codes.

Appendix 7B Motorized swivel

30^{o}

Appendix 7C Operationalization of experiment

Figure 7C1 shows the procedure of the experiment. The experiment was supervised by two persons. One person was the experimental leader. The experimental leader welcomed the participant supplying the tasks and questionnaires to the participant. The experimental leader was the only one having contact with the participant. There was also a technical leader. The technical leader was in the observation room and controlled the cameras, caused the failures and observed the participant trough the mirror (between the living room and the observation room). The technical leader did not make contact with the participant and was invisible for the participant. The activities presented in figure 7C1 will be explained below.

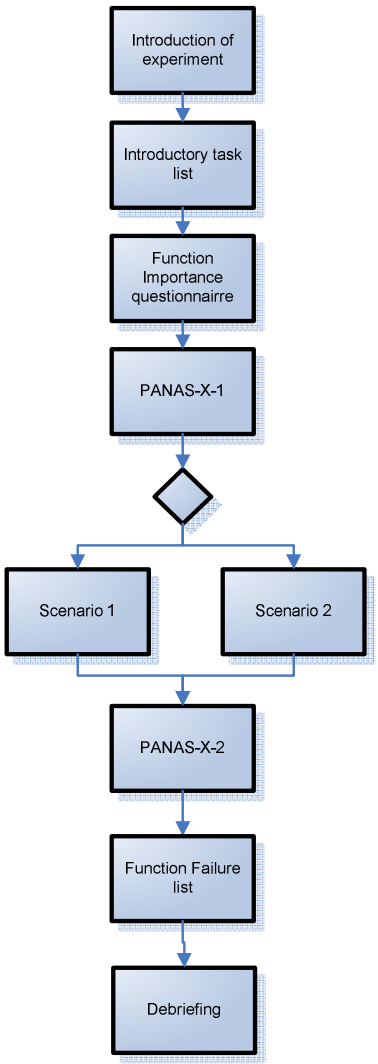


Figure 7C1: Design of the experiment

Introduction of experiment

The participant was welcomed, offered a drink and guided to the living room. The participant was told that he/she is going to test a television on user friendliness and that it is the television which is at interest

and not the participant itself. The participant was informed that this experiment will be filmed and that he had to do a couple of tasks.

Introductory task list

To make the participant comfortable with the situation and the television, the participant first had to do an introductory task. The participant got a list where a couple of functions were explained, like dual screen and the motorized swivel. An exploratory task list for being familiar with these functions had to be completed by the participant. The acquaintance of all participants with the television used in this experiment was set with this task list to the same level. The observer stayed during this introduction in the living room. The participant was also told that he/she can ask questions when things are unclear.

Function Importance questionnaire

The experiment is continued with filling in a function importance questionnaire. As said above, participants could rate the function of the motorized swivel as more important after using it. This could result in a situation that there are no homogeneous groups anymore. The results of the function importance questionnaires were used to decide which scenario would be executed by the participant. A low function importance score on the item: use of the motorized swivel (in comparison with the item: watch the desired program without failures in screen and sound) resulted in a continuation with scenario 1. Otherwise the participant had to continue with scenario 2. (See for a more detailed discussion Gielen, 2007). With this procedure two homogenous groups were created.

PANAS-X-1

Before the participant continued with one of the two scenarios he first had to complete a PANAS-X questionnaire. This PANAS-X questionnaire was used to measure the base-situation for the UPFS measurement. In summary, participants had to fill in this PANAS-X questionnaire with regard to what extent they felt right now, that is, at the present moment. This PANAS-X questionnaire is called PANAS-X-1 in the remainder of this master thesis.

Scenario 1 and Scenario 2

As said above, based on the results of the function importance questionnaire a decision will be made about how to continue with the experiment. A participant could continue with scenario 1 or scenario 2. Scenario 1 and 2 will be described below:

Scenario 1: UPFS of a low importance function

The first scenario was developed for testing how participants perceived the failure of a low importance function. This scenario started with watching television: start at channel one and go through the other channels to see what is broadcasted. When the participant found a program he/she liked he/she should watch this program for a while. Then the participant should take place behind the laptop (see figure 6.2.1). At the laptop there was a task which requires to fill in which program is broadcasted at which channel. The participant cannot watch the television from this point and so should turn the LCD-TV with the motorized swivel. After completion of this task he/she should go back to the couch and watch television again. At the couch it was not very easy and comfortable to watch television when the LCD TV was turned to the laptop. So, the participant should turn back the LCD TV with the motorized swivel. At this time the technical leader caused a failure in the motorized swivel. The LCD TV would not turn. This failure endured for 2 minutes. The motorized swivel “suddenly” (after 2 minutes) worked again. On this point there were two possible states in which the experiment could be:

State 1: The participant was still trying to use the motorized swivel.
State 2: The participant was not trying to use the motorized swivel anymore.

The difference between state 1 and state 2 is that the participants in state 1 would experience that the motorized swivel would work again. The participants in state 2 would not experience a "working again" motorized swivel. This results in a situation that the participants are not completely homogenous anymore. Therefore, when the participant gave up his/her attempt to use the swivel before the two minutes have passed the experimental leader will return to the living room and asked if everything went well. The participant would probably notice that the swivel did not work (anymore). The experimental leader asked the participant to show this. Then the participant would see that the motorized swivel worked (again). In the case that the participant did not tell the experimental leader that there was a problem with the swivel[6], the experimental leader self would turn the LCD screen with the remote control. During this scenario the experimental leader was not in the living room.

Scenario 2: UPFS of a high importance function

The second scenario was developed for testing how participants perceived the failure of a high importance function. In general, this task could be described as: watch television, start at channel one and go through the other channels to see what is broadcasted. When the participant found a program he/she likes, he/she should watch this program for a while. The participant should also try different functions of the television. During this watching of television at a certain moment of time a failure occurred. The technical leader could disturb the broadcast-signal. This disturbance will last for two minutes then the signal will be set to normal. The participant would (always) experience that the picture is normal again. During this scenario the experimental leader was not in the living room.

PANAS-X-2

Right after the completion of the scenario the experimental leader asked the participant to fill in PANAS-X again. Participants had to fill in this PANAS-X questionnaire with regard to what extent they felt right now, that is, at the present moment. This PANAS-X questionnaire is called PANAS-X-2. The first score on PANAS-X (PANAS-X-1) was compared with the second score on PANAS-X (PANAS-X-2). The difference is a measure for UPFS.

Function Failure list

After the completion of PANAS-X-2 the participant got a list of functions of the LCD-TV (function failure list) and the participant should indicate if this function worked flawless or not (see figure 7C2 below). This function failure list is used to check if the failure impact is equal between both scenarios. In figure 2.1 is presented that failure impact influences UPFS. Within this experiment the influence of failure impact is not examined. However it is necessary to control for failure impact. With this function failure list it is examined if controlling for failure impact succeeded. When the participant indicated that a function contained a failure, he/she should give score between 0-5. A zero means that the function could be used very well (despite the failure) and a five means that the function almost could not be used. The score should be equal for both scenarios. This score is called the function impact severity in the remainder of this thesis.

[6] It is possible that the participant does not dare to say that the motorized swivel does not work, because the participant might think that he/she has caused the failure.

Hieronder staat een tabel met functies die de door jou geteste TV bezit. Geef voor al de door jou geteste functies aan of ze naar jouw mening foutloos werkte (zet dan een kruisje in het vakje foutloos) óf juist faalde tijdens het gebruik van de TV. Indien jij van mening bent dat (een) bepaalde functie(s) faalde(n), kun je dan een score tussen de 0-5 geven voor de ernst van dit falen. Een 0 score betekent dat de functie nog goed te gebruiken was en een 5 score betekent dat de functie bijna onbruikbaar was geworden. Als je een bepaalde functie niet hebt gebruikt kruis dan NVT aan.

Nr.	Functies	Foutloos	Fout, Score:	NVT
1.	**Kijken van het gewenste programma** Kijk het gewenste programma op TV, zonder mankementen in beeld en geluid.			
2.	**Gebruik van teletekst** Gebruik teletekst en alle functies die ertoe behoren.			
3.	**Kijk twee programma's tegelijk** Kijk twee programma's tegelijk met Dual Screen. In het hoofdscherm wordt een subscherm opgeroepen waardoor je naar twee tv-zenders tegelijk kunt kijken, of naar één zender en teletekst.			
4.	**Mute** Schakel het geluid van de TV uit, met één druk op de knop.			
5.	**Veranderen van het beeldformaat** Selecteer het gewenste beeldformaat. Verschillende beeldformaten zijn: Automatisch, Super zoom, 4:3, 14:9, 16:9, Ondertitel zoom en Breedbeeld.			
6.	**Vorige Tv-zender** Schakel tussen de Tv-zender waar u nu naar kijkt en de zender waar u het laatst naar gekeken hebt door één druk op de knop.			
7.	**Gebruik van de gemotoriseerde draaivoet** Gebruik de gemotoriseerde draaivoet om het scherm te draaien tussen + en – 30 graden, voor een optimale kijkhoek.			

Figure 7C2: Function Failure List

Debriefing

After completion of this questionnaire the participant was debriefed and thanked for participation. The participant got a little present. The participant was told the real goal of the experiment and asked not to pass this on.

Appendix 7D Redesign of experiment

- The scenarios had to be more directive.
  With regard to scenario one: A small adaptation is made in the scenario. First it is explicitly stated that the participant should take along the remote control when taking place behind the laptop. This will emphasize the idea that they should turn the LCD screen through the motorized swivel by using the remote control. Another adaptation is that it is explicitly stated that the participant should turn back the screen when he/she completed the task on the laptop and is going to sit at the couch.
  With regard to scenario two: Participants did not really notice the difference between normal picture and a bad picture. Participants thought that it was not a failure of the LCD-TV, but that this bad picture was caused by the cable company. So, the participants did not perceive this as a failure. This scenario had been changed into the task that the participant first had to start at channel one and go through the other channels to see what is broadcasted. Next they had to use some other functions of the television which are seen as interesting by participant. Subsequently the participant had to fill in a form at the couch about what kinds of programs were broadcasted at the channels. During this task a failure occurs at the picture. After three minutes the picture is set to normal. With this procedure it is ensured that the participant had seen every relevant channel before the failure in a normal condition. The participant should notice that during the failure the picture is very bad or at least worse than in the normal condition.

- The function failure list had to be shorter.
  Some of the functions stated at the list were not tested/used by the participant. These functions were removed (like automatic channel programming, channel programming by hand, adding preferences in sound, picture etc., switch on timer, sleep timer, and child lock)

- PANAS-X could not measure the caused irritation as expected
  The results showed that the negative emotions didn't worsen due to the failures occurred during the experiment. Apparently the failure didn't have enough impact on the participant. The failure occurrence had to be longer. So, the failure duration is set from 2 to 3 minutes. For measuring UPFS another questionnaire (exit questionnaire[7]) was added (see figure 7D1). This question contained 10 propositions about irritation and action. After completion of the function failure list the exit questionnaire had to be filled in. Normally should appear from the function failure list that one function failed during the experiment (the picture or the motorized swivel). The exit questionnaire had to be filled as if the participant had bought this television a couple of weeks ago and "the participant's" television had the failure as noticed before. In the case the participant did not write down any failure or more than one failure on the function failure list, the experiment leader will tell the participant to fill in this questionnaire as if the television had a picture failure or a failure in the motorized swivel. The choice depended on the actual scenario the participant went trough.

[7] There will be returned to this exit questionnaire later on.

Stel je nu de situatie voor waarin je deze televisie hebt gekocht. Hij staat sinds een paar dagen bij jou in de huiskamer. Vervolgens treden de problemen op in de televisie die jij net ook opmerkte. Hoe zou jij reageren?

		Zeer waarschijnlijk				**Zeer onwaarschijnlijk**
		1	**2**	**3**	**4**	**5**
1.	**Ik zou deze fout enorm irritant vinden.**	O	O	O	O	O
2.	**Ik zou de klantenservice van Philips raadplegen voor hulp.**	O	O	O	O	O
3.	**Ik zou me behoorlijk opwinden over dit probleem.**	O	O	O	O	O
4.	**Ik zou gewoon wachten totdat het probleem verdwenen was.**	O	O	O	O	O
5.	**Over een dergelijke fout zou ik zeker niet boos worden.**	O	O	O	O	O
6.	**Deze televisie zou ik direct terugbrengen naar de winkel!**	O	O	O	O	O
7.	**Ik zou op het Internet gaan zoeken naar mogelijke oplossingen voor dit probleem.**	O	O	O	O	O
8.	**Thuis zou ik deze fout beschouwen als een kleine tekortkoming in de televisie**	O	O	O	O	O
9.	**Als dit probleem thuis op zou treden zou ik een familielid/kennis om hulp vragen.**	O	O	O	O	O
10.	**Aan een dergelijk probleem zou ik me niet ergeren...**	O	O	O	O	O

Figure 7D1: Exit questionnaire

Appendix 8A PANAS-X results UPFS-score

	Fear*	Hostility*	Guilt	Sadness	Basic Negative Emotion*	General Negative Affect*
Scenario 1 (N=12)	1.83	1.83	1.33	0.33	2.75	1.33
Scenario 2 (N=13)	0.00	-0.69	0.38	-0.85	0.15	-0.29

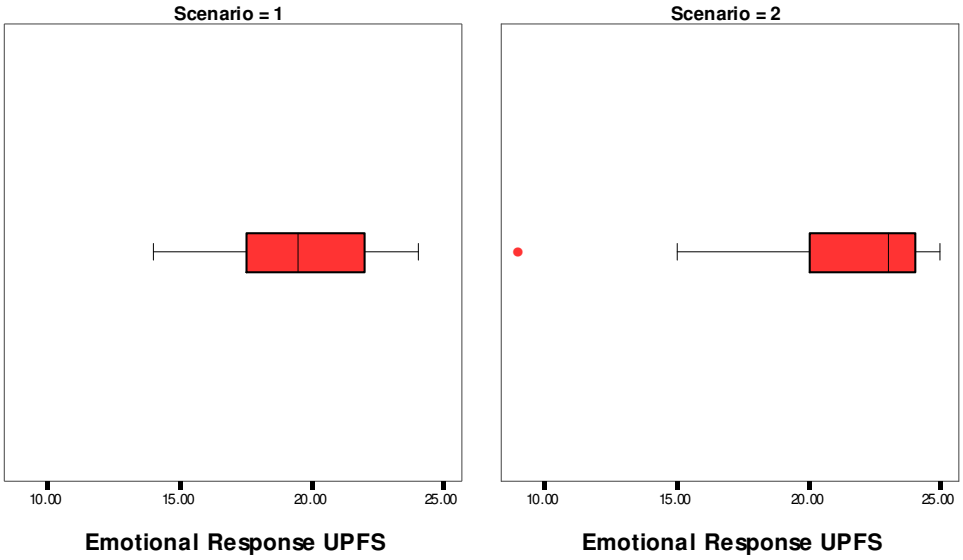
* = significant different between scenario 1 and scenario 2

Appendix 8B Exit questionnaire

Item	Response*	Abbrevation
I would find this failure enormously irritating	ER	enormous irritating
I would seek advice at the customer service of the manufacturer	AR	go to service-desk
I would be quite enraged about this problem	ER	become enraged
I would just wait till the problem disappeared	AR	just wait
I would not get angry about such a failure	ER	not angry
I would immediately return this TV to the store	AR	return tv back
I would have a look on the internet to find a solution to this problem	AR	searching at Internet
At home, I would consider this a small demerit of the TV	ER	small failure
If this problem occurs at home I would ask a relative/friend for help	AR	ask relative for help
I would not annoy myself about such a problem	ER	does not bother me

*ER = Emotional Response
AR = Action Response

Appendix 8C Box plot exit questionnaire scenario 1 and 2



Appendix 8D Categories of outliers and their sources

Category	Source
Procedural error	Wrong data entry or a mistake in coding.
Extraordinary event	For example, tracking the average daily rainfall, when there was a hurricane that lasts for several days and record extremely high rainfall levels.
Extraordinary observations	For this class of outliers the researchers has no explanation. In these instances a unique and markedly different profile emerges.
Unique in their combination	In this case the observations fall within the ordinary range of values on each of the variables. These observations are not particularly high or low, but are unique in their combination of values across the variables.

Source: Hair et al. 2005

Appendix 8E Effect size

An indication for the expected effect size can be calculated as follows. Using the results of table 8.5: the mean difference between the UPFS of scenario 1 and scenario 2 is 22.08 – 19.50 = 2.58. The standard deviation of scenario 1 is 3.26 and of scenario 2: 3.06. The effect size can be calculated, using Cohen's *f:*

f = effect size

$$f = \frac{\sigma_\mu}{\sigma}$$

where

$$\sigma_\mu^2 = \frac{\sum_{i=1}^{k} n_k \left(\mu_i - \mu\right)^2}{N}$$

σ_μ = standard deviation of the effect

σ = standard deviation in the population

k = number of groups

n_i = sample size of group i

μ_i = population mean in group i

μ = population mean of the total sample size

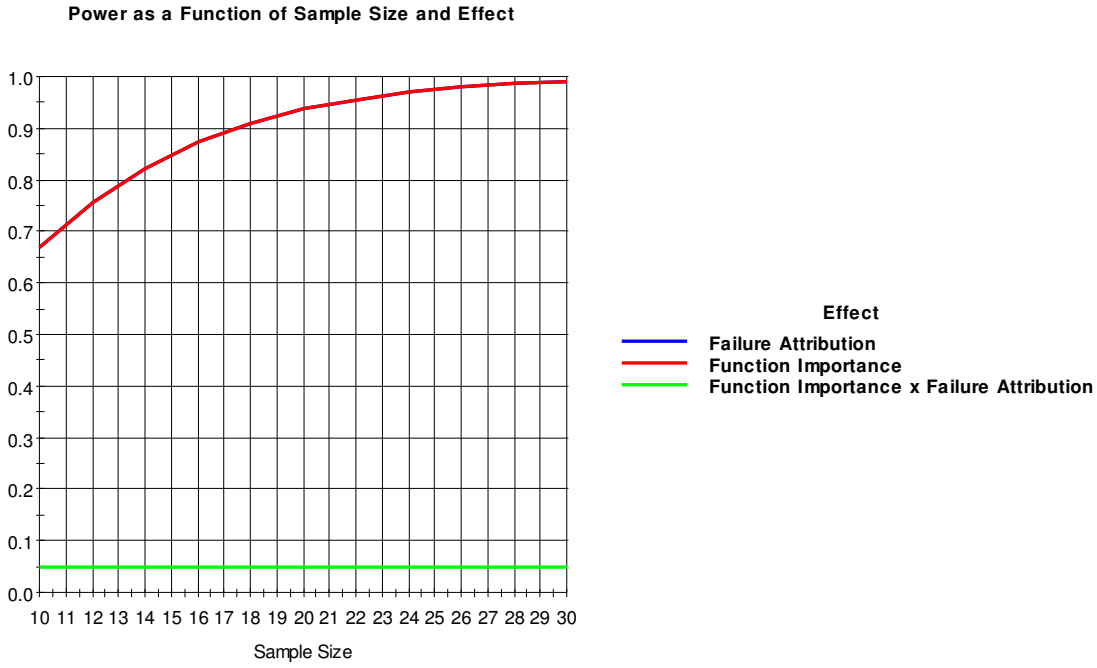
N = total sample size

$$\sigma_\mu^2 = \frac{12*\left(19.5-20.79\right)^2 + 12*\left(22.08-20.79\right)^2}{24} = 1.66$$

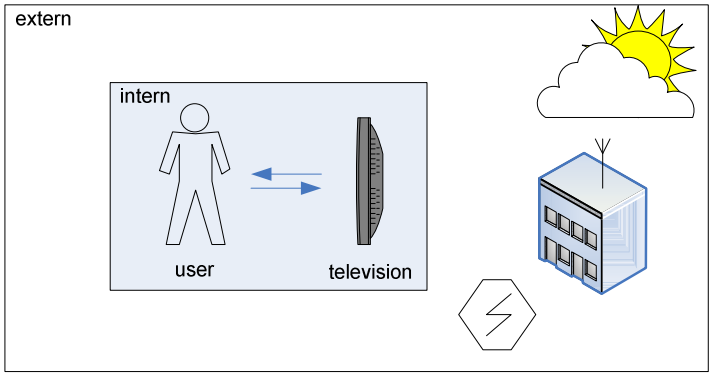
$$f = \frac{\sqrt{1.66}}{3.36} = 0.38$$

Appendix 8F Power as a function of sample size (per group) and effect

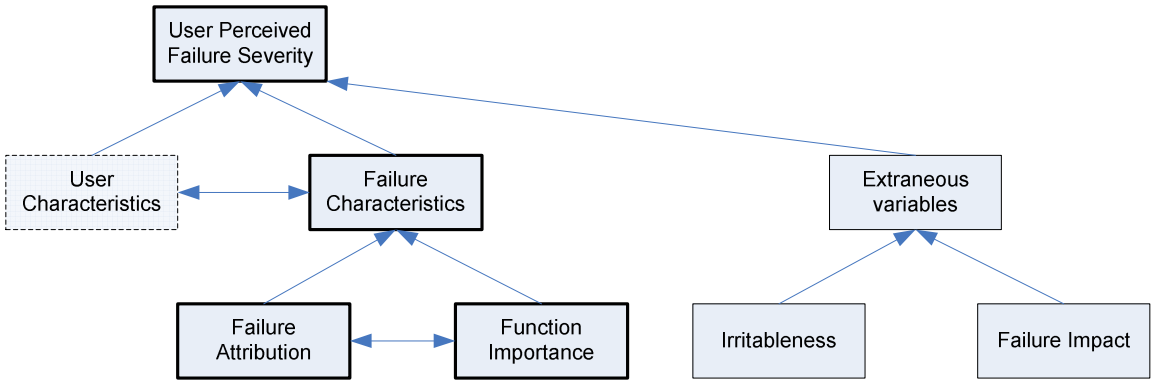
The line of failure attribution and function importance overlap.



Appendix 9A Intern and extern failure attribution



Appendix 9B The revised research model



Appendix 10A Design of experiment 2

Figure 10.A presented on the next page gives a flowchart of the experiment. The various blocks of this flowchart will be explained below.

Introduction of experiment
The participant will be welcomed and told that he/she is going to test a television on user friendliness. It is explicitly told that the television which is it at interest and not the participant itself. Potential participants are stimulated to participate by raffling a LCD-screen among the participants.

PANAS
Participants had to fill in PANAS, a list of two 10-items mood scales: 10 items Positive Affect scale and 10 items Negative Affect scale. These two scales are the same scales as the General Dimension Negative Affect and the General Dimension Positive Affect of PANAS-X, Appendix 3C, figure 3C5. PANAS was used to control for irritableness afterwards. Afterwards, participants who had not an average equal level of irritableness were excluded from further analysis. In contrast with the previous experiment controlling for irritableness was done afterwards. Controlling afterwards was done, because approaching participants in advance with PANAS and than after the analysis making an appointment is more time-consuming for the participants than doing the experiment right away. The only disadvantage is that more participants are needed due to the selection afterwards.

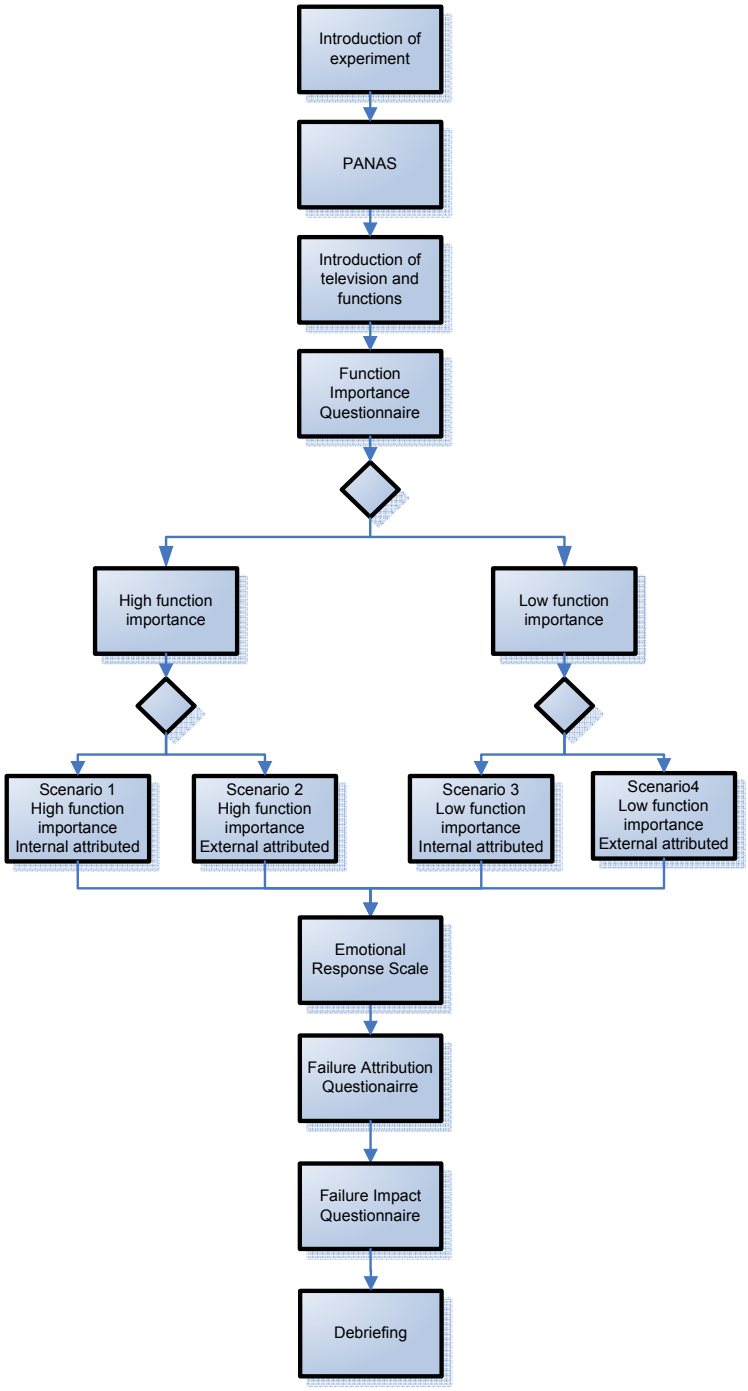


Figure 10A: design second experiment

Introduction of television and functions

A film was showed were the television was introduced and some particular functions were showed how they worked. The viewing quality, the use of the motorized swivel, ambilight and the use of dual screen was showed.

Function Importance Questionnaire

As argued in the previous experiment participants could rate the function importance of the motorized swivel different after using it. The results of the function importance questionnaires were used to decide

which scenario would be executed by the participant. A low function importance score on the item: use of the motorized swivel (in comparison with the item: watch the desired program without failures in screen and sound) resulted in a continuation with the low function importance scenario. Otherwise the participant had to continue with the high importance scenario. (See for a more detailed discussion Gielen 2007). With this procedure two homogenous groups were created. This function importance questionnaire only contained the items: watching the desired program flawless, watching two programs at the same time (dual screen), ambilight, use of motorized swivel.

Scenarios
Based on the two independent variables of figure 9.2 there were four different scenarios developed. For each of these scenarios a different film was made. Within this film was shown that the particular function worked, subsequently that the particular function didn't work. Also the cause of this failure was made clear. This was done by a slideshow, text and pictures explicit pointing at the causes. For scenario 1: High function importance internal attributed, reason: a software failure occurred within the television which lead too no picture/or a very bad picture quality, television needs to be reset. For scenario 2: High function importance external attributed, reason: a disruption of the signal caused by the cable company. For scenario 3: Low function importance internal attributed, reason: also a software failure. For scenario 4: Low importance scenario external attributed, reason: a disruption in power supply.

The other extraneous variables are kept constant:

- Failure Impact: all failures had a 100% loss of functionality as a result
- Failure Frequency: in every scenario one failure occurred with the picture or with the motorized swivel.
- Failure Reproducibility: The failures in both functions were presented if they occur at random. When the participant had to fill in the Emotional Response scale he/she was told that he/she bought this television a couple of weeks ago and contained the failure which just occurred. So, it is made plausible that these failures occur at random.
- Failure solvability: The failure solvability is kept constant by stating that the problem will be solved in a half hour. The overheated swivel will cool down in a half hour. The cable disruption and peak-load in power supply will endure for a half hour. The software problem can be solved by resetting the television. This will be done automatically by the television. The television will re-install the software which will endure about a half hour.
- Failure work around: The participant is not able to evade the failure. Physically not, because only a film and slides are given about the failure so, the participant does not interact (physically) with the television. Imaginary also would be difficult because, causes are listed why the failure occurred and these causes are beyond the control of the user like: a software failure, cable signal or a disruption in power supply.
- Failure time: The participant was not confronted a ten minutes with a non working function. Instead the participant was given that this function will not work for 10 minutes.
- Controllability: All occurred failures are not under control of the participants and were equal for the four different scenarios
- Stability: This variable was related to failure reproducibility. The stability of the failure was the same for all four dimensions. The failures were presented if they occur at random and it was not clear to the participant if this failure could happen again in the future.

Emotional Response Scale
After the film/slides, where the failure and its causes were showed, the participant had to fill in the Emotional Response Scale. This scale was used for measuring UPFS. The emotional response scale is given at Appendix 7D, Figure 7D1**.**

Failure Attribution Questionnaire
A failure attribution questionnaire will be added for controlling if the participant also perceived that the failure is occurred internally (the interaction between user and television) or externally (factors beyond the interaction between user and television), see also figure 9.1. This questionnaire is based on Russell's Causal Dimensions Scale. This scale has been well received in literature (Russell 1982). The scale is internally consistent and has a reliability of $\alpha = 0.867$. For this master thesis only the locus dimensions is at interest. This dimension consists of three items and the Dutch version is given in Appendix 10B.

Failure Impact Questionnaire
The failure impact questionnaire will be added to control for failure impact. This questionnaire will contain only one item. This item will be the particular failed function: watching the desired program or use of the motorized swivel. The Dutch version of this questionnaire is given below:

Failure Impact Questionnaire

Motorized Swivel

In de zojuist getoonde presentatie/film is een fout getoond in een van de functies van de televisie. Wat is de impact van deze fout volgens jou? Geef een score van 0 tot 5. Een nul score betekend dat ondanks de fout de functie nog zeer goed gebruikt kan worden. Een vijf betekend dat de functie bijna niet meer gebruikt kan worden. De functie van de televisie waar de fout optrad is hieronder gegeven, omcirkel aub de impact van de falende functie:

De gemotoriseerde draaivoet
Het instellen van de gewenste kijkhoek was:

Foutloos 0 1 2 3 4 5 bijna niet bruikbaar

Picture

In de zojuist getoonde presentatie/film is een fout getoond in een van de functies van de televisie. Wat is de impact van deze fout volgens jou? Geef een score van 0 tot 5. Een nul score betekend dat ondanks de fout de functie nog zeer goed gebruikt kan worden. Een vijf betekend dat de functie bijna niet meer gebruikt kan worden. De functie van de televisie waar de fout optrad is hieronder gegeven, omcirkel aub de impact van de falende functie:

Het kijken van het gewenste tv-programma
Kijken van het gewenste tv-programma zonder mankementen in beeld en geluid was:

Foutloos 0 1 2 3 4 5 bijna niet bruikbaar

Debriefing
After completion of the failure impact questionnaire the participant was debriefed and thanked for participation. The participant was told the real goal of the experiment and asked not to pass this on. The participant was contacted about the result of the lottery.

All these questionnaires and video films were processed in an automated design. When the participant starts with the experiment, no further intervention or comments of the experimental leader were necessary. However, when the participant has some questions he/she can always ask these questions.

Appendix 10B Failure Attribution Questionnaire

Denk aan de oorzaak van de zojuist getoonde fout. Wie denk je dat verantwoordelijk is voor het ontstaan van de zojuist getoonde fout? Hieronder staan steeds twee tegengestelde stellingen. Vink de cirkel aan die het beste de juiste situatie weergeeft.
Bijvoorbeeld: Ben je het geheel eens met de linker stelling vink dan de linker cirkel aan. Ben je het enigszins eens met de rechter stelling vink dan de tweede cirkel van rechts aan.

De fout gebeurde door iets wat de televisie deed	○ ○ ○ ○ ○	De fout geeft een aspect weer van de prestatiekwaliteit van andere dingen
De fout gebeurde door iets wat andere dingen of personen deden	○ ○ ○ ○ ○	De fout werd veroorzaakt door een onderdeel van de televisie
De fout is een aspect van de prestatiekwaliteit van de televisie	○ ○ ○ ○ ○	De fout werd veroorzaakt door iets anders

Appendix 10C Revised Failure Attribution Questionnaire

Denk aan de oorzaak van de zojuist getoonde fout. Wie denk je dat verantwoordelijk is voor het ontstaan van de zojuist getoonde fout? Hieronder staan steeds twee tegengestelde stellingen. Vink de cirkel aan die het beste de juiste situatie weergeeft.
Bijvoorbeeld: Ben je het geheel eens met de linker stelling vink dan de linker cirkel aan. Ben je het enigszins eens met de rechter stelling vink dan de tweede cirkel van rechts aan.

De fout gebeurde door iets wat de televisie deed	○ ○ ○ ○ ○	De fout gebeurde door iets wat andere dingen of personen deden
De fout is een tekortkoming in de prestatiekwaliteit van de televisie	○ ○ ○ ○ ○	De fout is een tekortkoming in de prestatiekwaliteit van andere dingen
De fout werd veroorzaakt door de televisie	○ ○ ○ ○ ○	De fout werd veroorzaakt door iets anders

Appendix 11A Number of participants in each scenario

User Perceived Failure Severity

N = 149

Function Importance	Failure attribution: intern	Failure attribution: extern
high	Scenario 1 Highest N = 35	Scenario 3 Low N = 38
low	Scenario 2 Higher N = 37	Scenario 4 Lowest ? N = 39

Appendix 11B Revised failure Impact Questionnaire

Motorized Swivel

In de tweede film die je zonet zag, faalde de “gemotoriseerde draaivoet”van de televisie. In hoeverre wordt volgens jou de bruikbaarheid van deze functie (de “gemotoriseerde draaivoet”) aangetast door dit specifieke probleem?

Kies een van de volgende opties:

Gemotoriseerde draaivoet
De functie, de “gemotoriseerde draaivoet” was in het **tweede** filmpje:

- ○ Goed bruikbaar
- ○ Redelijk bruikbaar
- ○ Enigszins bruikbaar
- ○ Niet bruikbaar
- ○ Bijna niet bruikbaar

Picture

In de tweede film die je zonet zag, faalde de “het kijken van het gewenste programma” van de televisie. In hoeverre wordt volgens jou de bruikbaarheid van deze functie (“het kijken van het gewenste tv-programma”) aangetast door dit specifieke probleem?

Kies een van de volgende opties:

Het kijken van het gewenste programma
De functie, “het kijken van het gewenste tv-programma” was in het **tweede** filmpje:

- ○ Goed bruikbaar
- ○ Redelijk bruikbaar
- ○ Enigszins bruikbaar
- ○ Niet bruikbaar
- ○ Bijna niet bruikbaar

Appendix 11C Assumptions underlying ANOVA

The assumptions underlying ANOVA are (Hair et al. 2005), (Green and Salkind 2002):

Independence of observations
A critical assumption is that the responses in each cell (group) are made independently of responses in any other group. No test can provide an absolute certainty of detecting all forms of dependence (Hair et al. 2005). All participants had to do the experiment under the same conditions.

Homoscedasticity
A second, (critical) assumption concerns the homogeneity of the variance-covariance matrices among the groups. There have to be no substantial differences in the amount of variance of one group versus another for the dependent variable. This can be checked with a Levene's test. When this assumption is violated, the p value and the resulting F test is untrustworthy. In the case this assumption is not met, it is preferred to use statistics that do no assume equality of population variances such as the Browne-Forsythe or the Welch statistic (Green and Salkind, 2002). When doing a post hoc test -in this case of not meeting this assumption- also tests that do not assume equal variances should be used like Dunnett's C (Green and Salkind, 2002).

Normality of the dependent variable
This assumption requires that the population distributions on the dependent variable to be normally distributed for all cells (groups). In a moderate (or larger) sample size, ANOVA may yield relatively accurate p values even when the normality assumption is violated (substantially). The power of the ANOVA-test may reduce considerably if the population distributions are nonnormal, and more specifically thick-tailed or heavily skewed. Testing for normality can be done with a Kolmogorov-Smirnov test or a Shapiro-Wilk test (in the case $N<50$). The significance of these tests are less useful in small samples (fewer than 30). Therefore it is recommended to use besides these tests also some graphical plots to assess the actual degree of departure from normality (Hair et al. 2005).

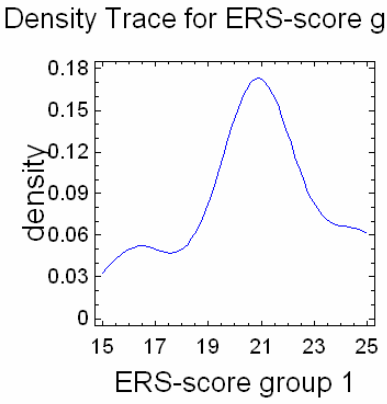
Appendix 11D Assumptions ANOVA for Irritableness

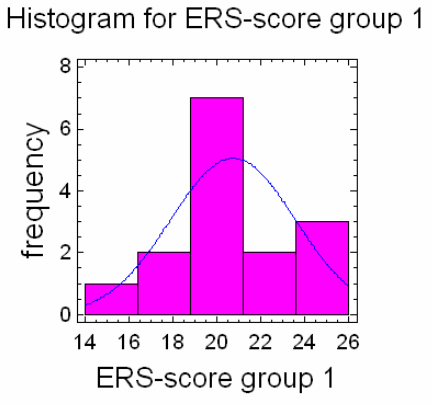
Independence of observations
The independence of the observations is ensured by the fact that all participants had to do the experiment under the same conditions.

Homoscedasticity
A Levene's test could not reject the hypothesis that the (error)variances are equal across the groups ($p = 0.14$).

Normality of the dependent variable ERS-score
The Shapiro-Wilk test could not reject ($p=0.32$) the hypothesis that the ERS-scores of group 1 are normally distributed. However this conclusion is based on N=15 and therefore less useful. A visual inspection of the histogram and density trace (see figure 11D1) of the ERS-score of group 1 shows that it is more likely to assume that the ERS- score of group 1 is not normally distributed.
The Kolmogorov-Smirnov test reject ($p \leq 0.01$) the hypothesis that the ERS-scores of group 2 are normally distributed. A visual inspection of the density trace and histogram (see figure 11D1) show that the distribution of the ERS-score of group 2 is somewhat negatively skewed.
The Shapiro-Wilk could not reject the hypothesis that the ERS-scores of group 3 are normally distributed ($p=0.08$). However this conclusion is based on N=21 and therefore less useful. A visual inspection of the histogram and density trace (see figure 11D1) of the ERS-score of group 3 shows also that the ERS- score of group 3 is not normally distributed.

There has to be noted that dividing the participants in three groups by setting threshold values affects the remaining distribution forms rigorously. Suppose that the ERS-scores of all participants (group 1,2 and 3) together are normally distributed. Cutting this group of ERS-scores in three parts by threshold values will have consequences for the distribution forms of the three individual parts. The distribution of the ERS-scores of Group 1 will lose his right tail. Group 2 will lose his left and right tail. Group 3 will lose his left tail. Therefore it is not strange that the ERS-scores of the three groups are not normally distributed.





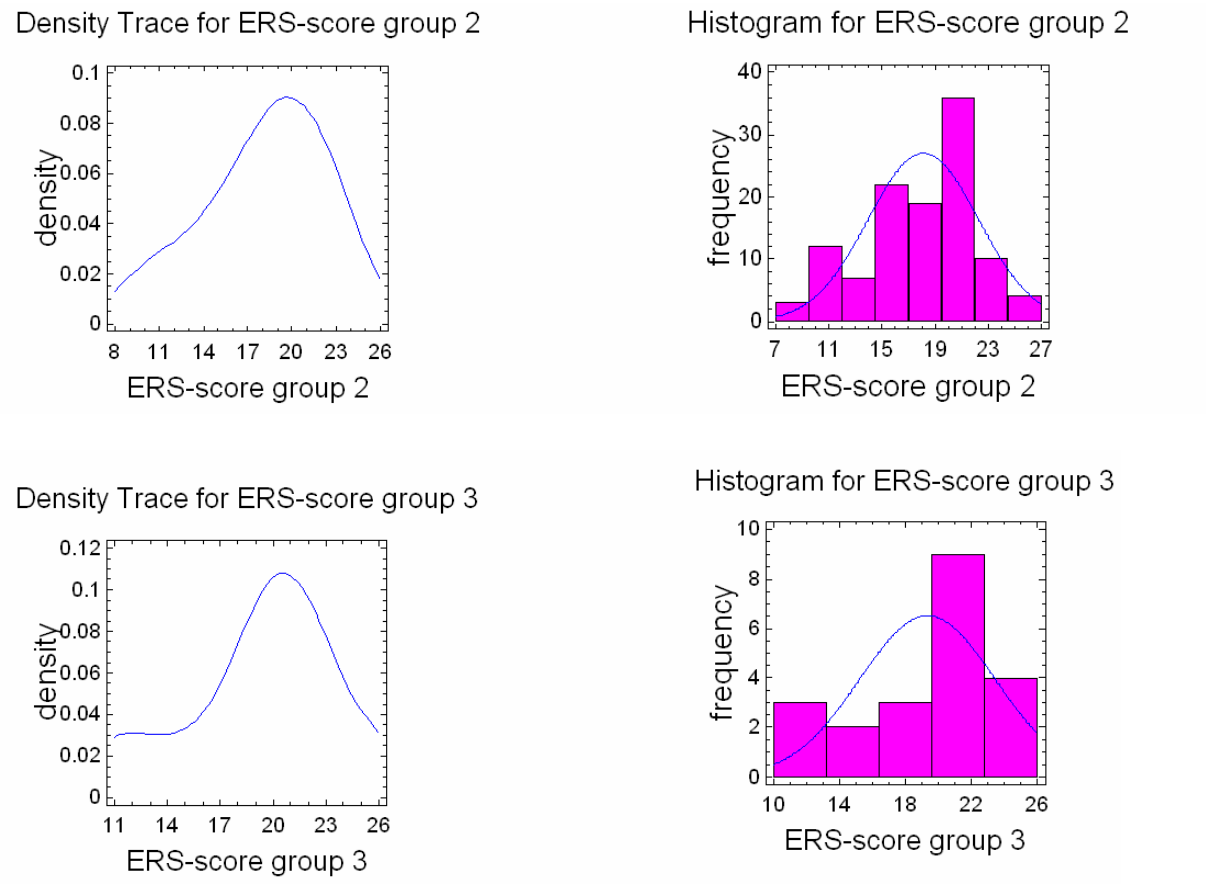


Figure 11D1: Density traces and histograms ERS-scores

Appendix 11E Assumptions ANOVA for Failure Impact

Independence of observations
The independence of the observations is ensured by the fact that all participants had to do the experiment under the same conditions.

Homoscedasticity
According to the result of the Levene's test the hypothesis is rejected that the (error)variances are equal among the groups ($p<0.05$). Therefore the Browne-Forsythe or the Welch statistic will be used instead of the F statistic.

Normality of the dependent variable FIQ
The normality assumption for Failure Impact Questionnaire is tested with a Shapiro-Wilk test. For all four scenarios the hypothesis that the FIQ-score is normally distributed is rejected, see table 11E1. Also the density traces and histograms confirm these results, see figure 11E1.

Table 11E1: Normality test for Failure Impact Questionnaire Score

	Shapiro-Wilk test	
Scenario*	N	p-value
1	25	0.003
2	27	0.000
3	26	0.005
4	31	0.007

Scenario 1: Failure Importance high, Failure Attribution intern
Scenario 2: Failure Importance high, Failure Attribution extern
Scenario 3: Failure Importance low, Failure Attribution intern
Scenario 4: Failure Importance low, Failure Attribution extern

It has to be noted that it is hard to get a normally distributed FIQ-score when the range of the FIQ-scores is limited between 1 and 5. In a moderate (or larger) sample size, ANOVA may yield relatively accurate p values even when the normality assumption is (substantially) violated.

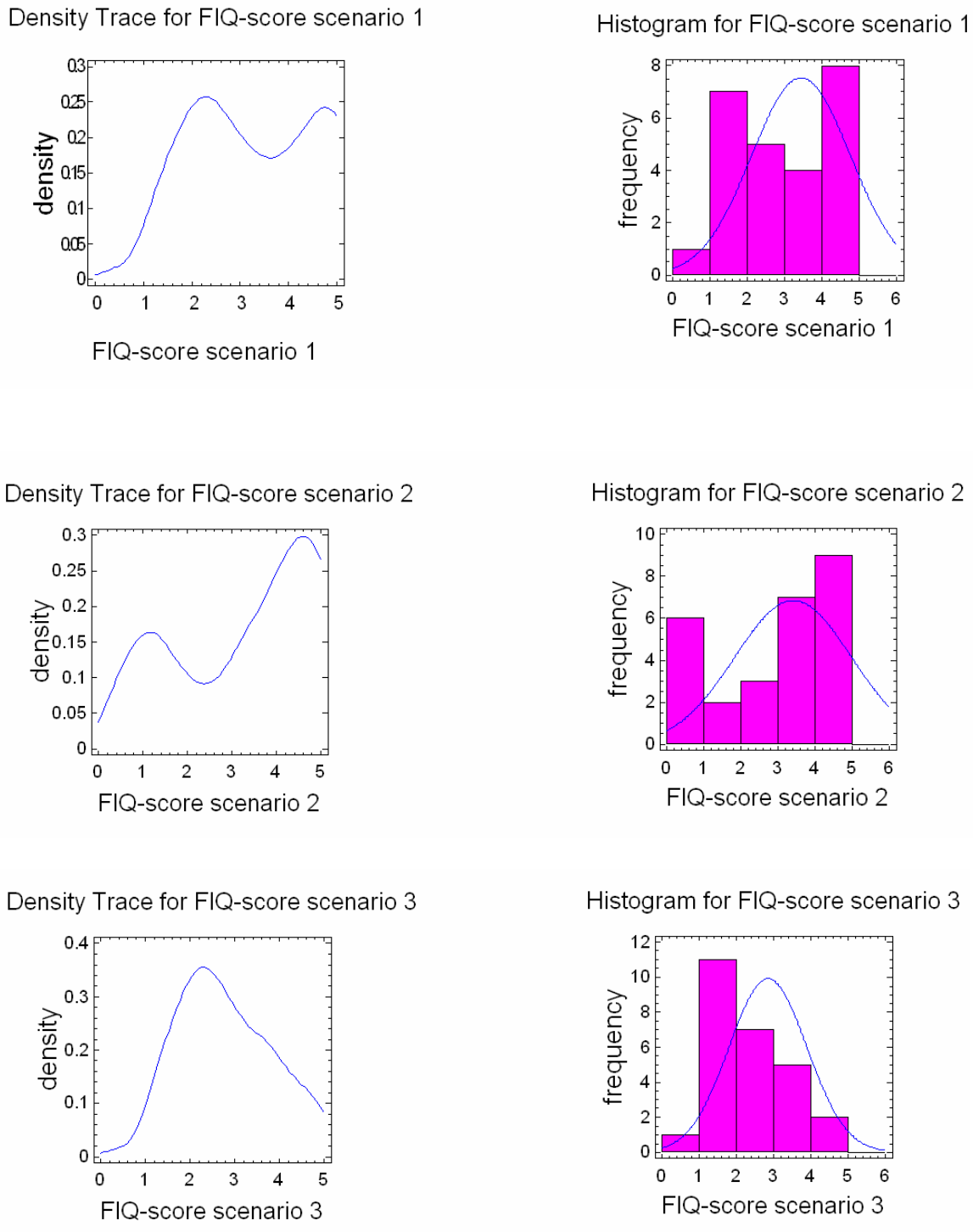


Figure 11E1: density traces and histograms

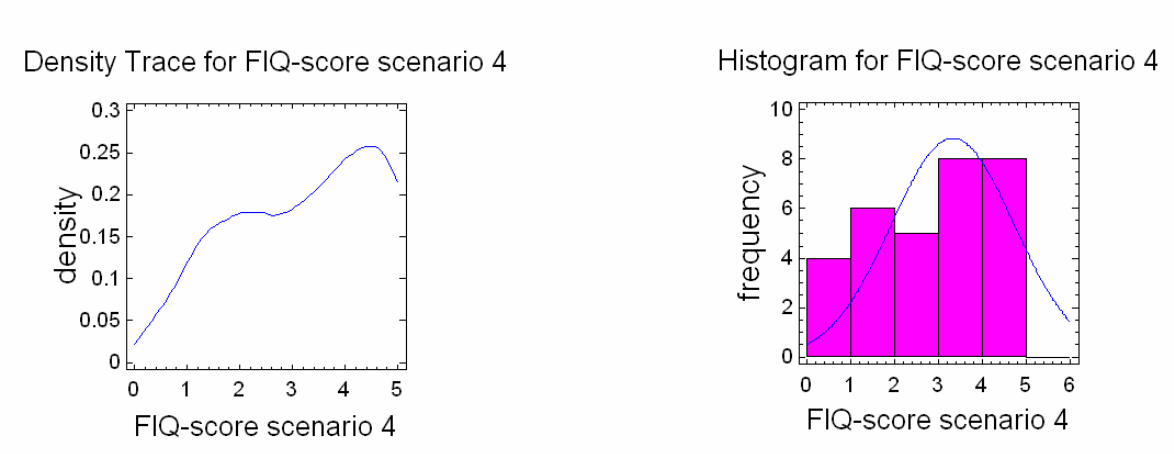


Figure 11E1 continued: density traces and histograms

Appendix 11F Assumptions ANOVA for Failure Attribution

Independence of observations
The independence of the observations is ensured by the fact that all participants had to do the experiment under the same conditions.

Homoscedasticity
The result of the Levene's test is marginally significant (p=0.05). The hypothesis, that the (error)variances are equal among the groups could not be rejected convincingly. However due to the marginal significance also the Browne-Forsythe statistic and the Welch statistic will be used for drawing conclusions.

Normality of the dependent variable
This assumption is not met. As expected a Kolmogorov-Smirnov test rejected the hypothesis that the distribution of a Failure Attribution Questionnaire-score is normally distributed. See also figure 11F1 for the accompanying density traces and histograms. However in a moderate (or larger) sample size, ANOVA may yield relatively accurate p values even when the normality assumption is (substantially) violated. The sample size is large per group (N=54 and N=59).

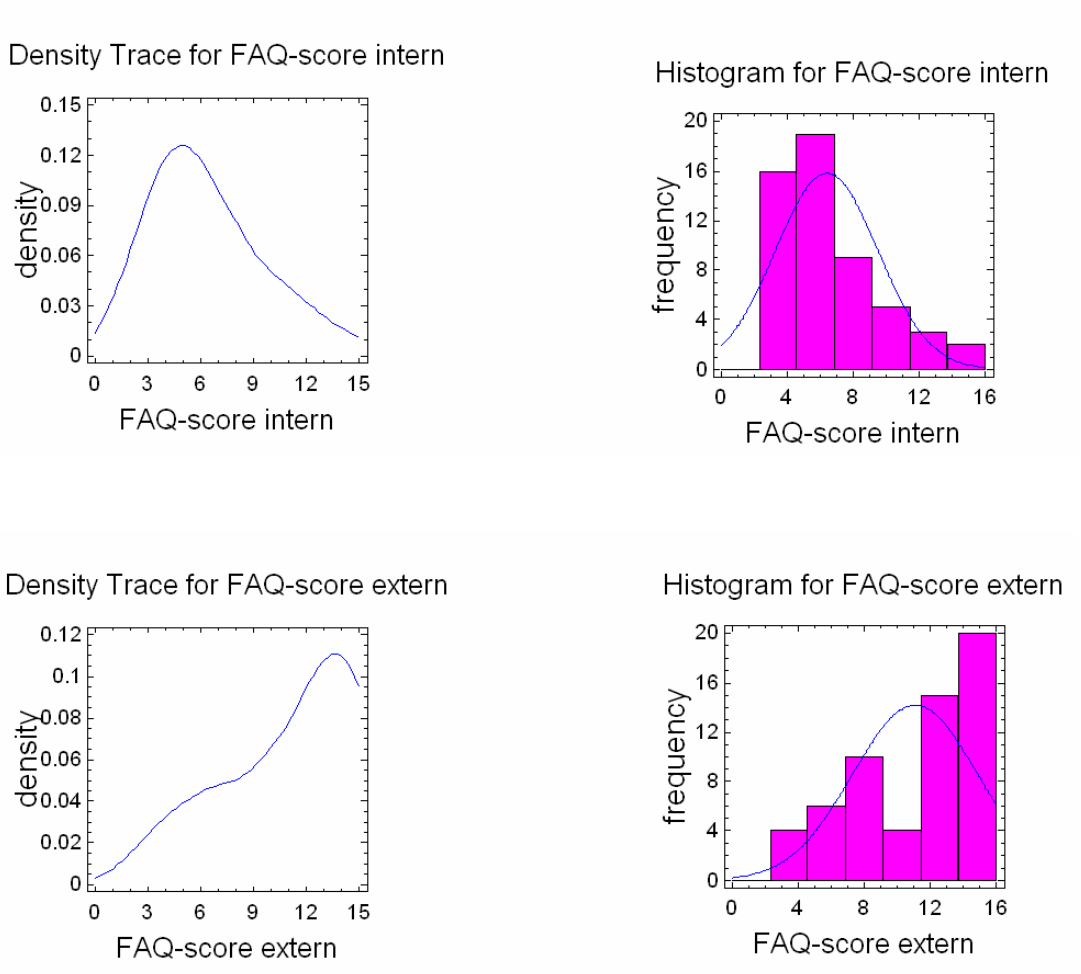


Figure 11F1: density traces and histograms

Appendix 11G Assumptions ANOVA for hypothesis testing

These hypotheses can be tested with a (two-way) ANOVA. The assumptions underlying ANOVA have to be checked first.

Independence of observations
The independence of the observations is ensured by the fact that all participants had to do the experiment under the same conditions.

Homoscedasticity
The result of the Levene's test is not significant (p=0.53) the hypothesis, that the (error)variances are equal among the groups could not be rejected.

Normality of the dependent variable ERS-score
This assumption requires that the population distributions of the dependent variable are normally distributed for all cells (groups). For all four scenarios a Shapiro-Wilk test was done. The results are presented in table 11G1. From these tests can be concluded that the ERS-scores of scenario 4 are not normally distributed see also figure 11G1. The ERS-scores from scenario 2 are just not normally distributed (p = 0.06). A detailed examination of the ERS-scores of these scenarios showed that there is no outlier in one of the four scenarios.

Table 11G1: Results Shapiro-Wilk test w.r.t. ERS-scores of the different scenarios

Scenario*	N	avg ERS	st.dev	Shapiro-Wilk test p-value
1	28	19.79	3.60	0.21
2	28	17.64	4.19	0.06
3	26	18.85	3.93	0.37
4	31	16.45	4.30	0.02

Scenario 1: Failure Importance high, Failure Attribution intern
Scenario 2: Failure Importance high, Failure Attribution extern
Scenario 3: Failure Importance low, Failure Attribution intern
Scenario 4: Failure Importance low, Failure Attribution extern

In the first three scenarios could the hypothesis that the ERS-scores are not normally distributed not be rejected. In scenario 4 this hypothesis is rejected. However in a moderate or larger sample size, ANOVA may yield relatively accurate p values even when the normality assumption is violated.

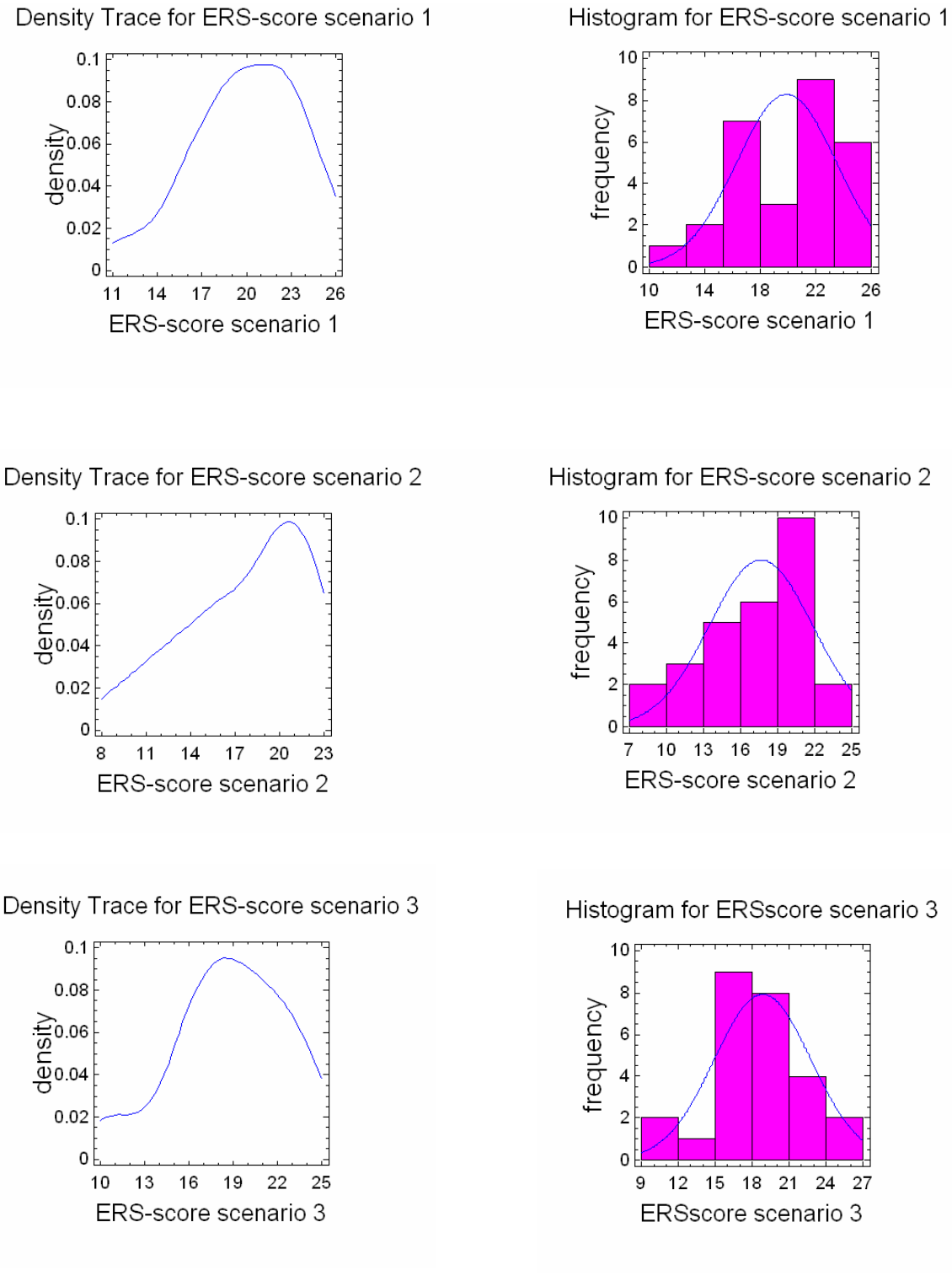


Figure 11G1: density traces and histograms

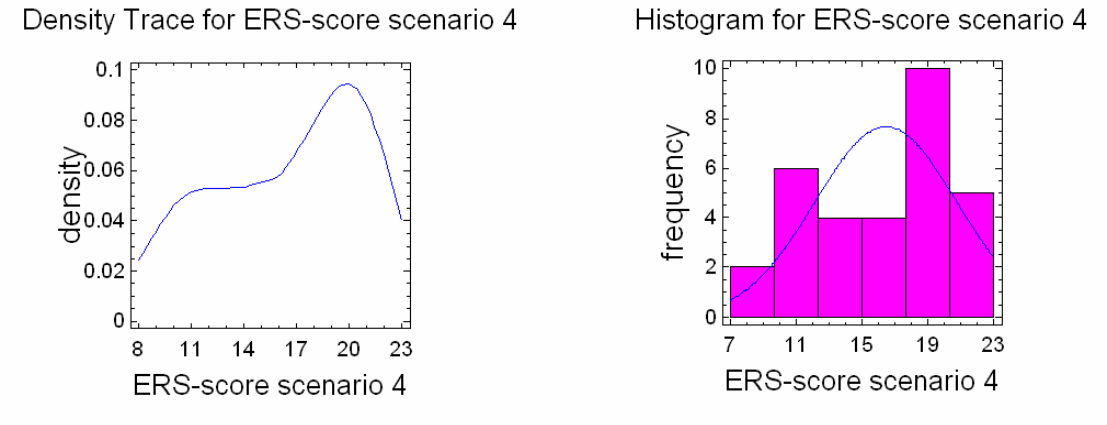
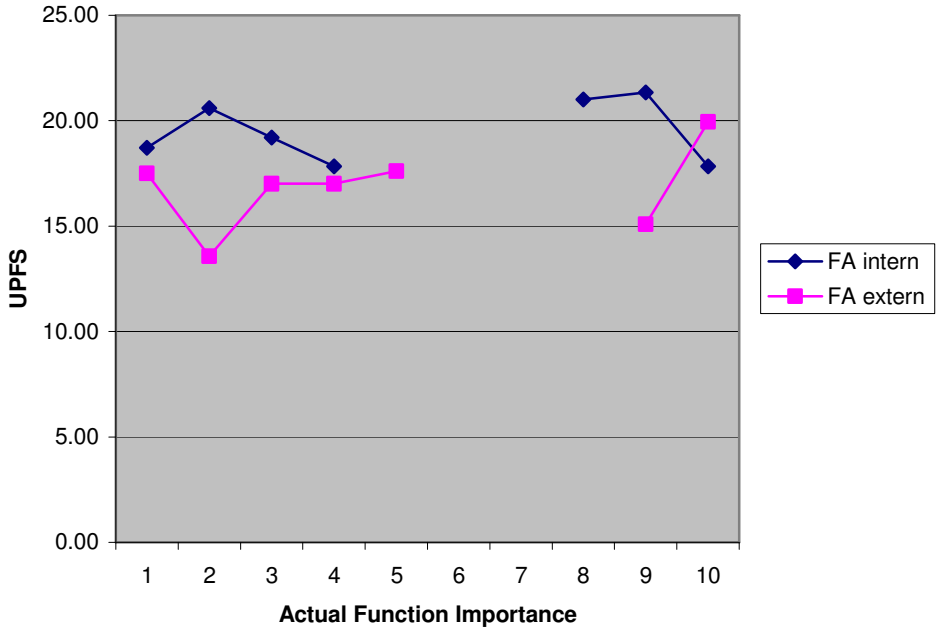


Figure 11G1 continued: density traces and histograms

Appendix 11H Relationship between FI, FA and UPFS

Cases with a sample size equal or smaller than 2 are not presented.



Appendix 11l Correlation matrix

Correlation matrix (Spearman's rho)

		ERS-score	Age	Negative Affect	Importance	Attribution	FIQ-score	FAQ-score
ERS-score	Spear. rho		-.08	.03	.13	-.26**	.33**	-.20*
	Significance		.41	.76	.19	.00	.00	.03
Age	Spear. rho	-.08		-.06	-.14	.08	.04	-.00
	Significance	.41		.54	.16	.43	.65	.99
Negative Affect	Spear. rho	.03	-.06		-.06	-.14	-.09	-.02
	Significance	.76	.54		.57	.15	.36	.88
Importance	Spear. rho	.13	-.14	-.06		-.01	.05	.36**
1 to 10	Significance	.19	.16	.57		.92	.62	.00
Attribution	Spear. rho	-.26**	.08	-.14	-.01		.09	.56**
1=intern 2=extern	Significance	.00	.43	.15	.92		.33	.00
FIQ-score	Spear. rho	.32**	.04	-.09	.05	.09		.04
	Significance	.00	.65	.36	.62	.33		.70
FAQ-score	Spear. rho	-.20*	-.00	-.02	-.36**	.56**	.04	
low = int high =ext	Significance	.03	.99	.88	.00	.00	.70	

** Correlation significant at the 0.01 level

* Correlation significant at the 0.05 level

Appendix 11J Relationship between function importance, failure impact and UPFS

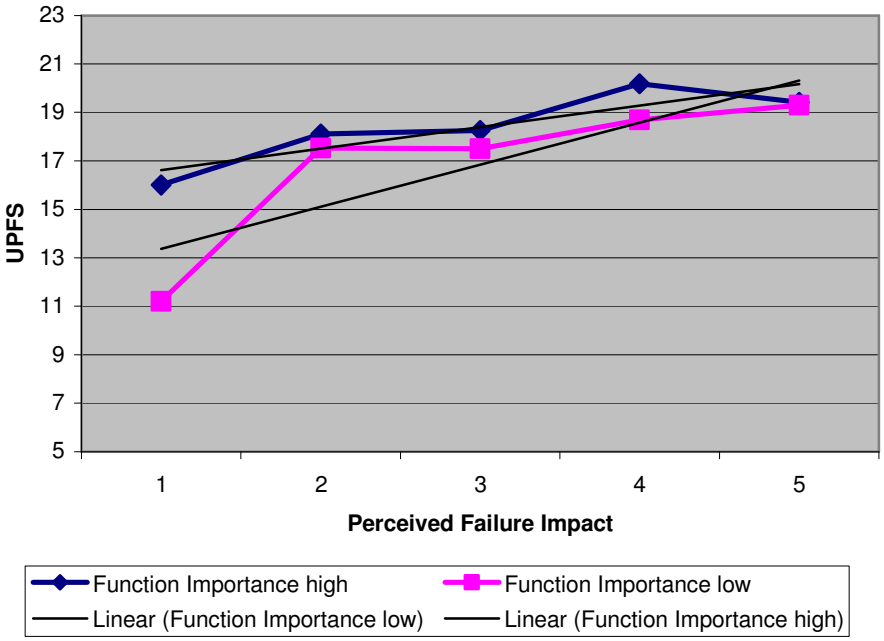


Figure 11J1: Relationship between function importance, failure impact and UPFS

In figure 11J1 UPFS is plotted against Perceived Failure Impact (FIQ-score) for Function Importance. A visual inspection reveals that an increase in Failure impact will lead to increase in UPFS. Also a possible interaction effect is visible. There are two linear trend lines added. These two lines cross at the end. However the graphs of low function importance and high function importance itself do not cross. The data behind this graph is presented in table 11J1.

Table 11J1: ERS scores for FI split for FIQ

FI*	FIQ-score*	ERS-score	N
High	1	16.00	7
High	2	18.11	9
High	3	18.25	8
High	4	20.18	11
High	5	19.41	17
Low	1	11.20	5
Low	2	17.53	17
Low	3	17.50	12
Low	4	18.69	13
Low	5	19.30	10

*FI = Function Importance
**FIQ= Failure Impact Questionnaire-score

Appendix 11K Stepwise regression, dependent variable ERS-score

			B	std err	Beta (In)	sig	VIF
step 1*	included	FIQ	1.04	0.29	0.33	0.00	1.00
	excluded	Age			(-0.11)	0.21	1.00
		NA			(0.09)	0.34	1.00
		FI			(-0.10)	0.28	1.01
		FA			(-0.32)	**0.00**	**1.00**
		FI*FA			(-0.26)	**0.00**	**1.00**
		FI*FIQ			(-0.06)	0.50	1.06
		FA*FIQ			(-0.37)	**0.00**	**1.32**
step 2**	included	FIQ	1.62	0.31	0.52	0.00	1.32
		FA*FIQ	-0.79	0.21	-0.37	0.00	1.32
	excluded	Age			(-0.13)	0.15	1.00
		NA			(0.07)	0.43	1.01
		FI			(-0.08)	0.39	1.02
		FA			(-0.13)	**0.56**	**7.09**
		FI*FA			(-0.11)	0.27	1.42
		FI*FIQ			(-0.03)	0.78	1.07

*Adj. $R^2 = 0.10$
**Adj. $R^2 = 0.20$

Appendix 11L Stepwise regression (blocks), dependent variable ERS-score

			B	std err	Beta (In)	sig	VIF
step 1*	included	FIQ	1.04	0.29	0.33	0.00	1.00
	excluded	Age			(-0.11)	0.21	1.00
		NA			(0.09)	0.34	1.00
		FI			(-0.10)	0.28	1.01
		FA			(-0.32)	**0.00**	**1.00**
		FI*FA			(-0.26)	0.00	1.00
		FI*FIQ			(-0.06)	0.50	1.06
		FA*FIQ			(-0.37)	**0.00**	**1.32**
step 2**	included	FIQ	1.13	0.27	0.36	0.00	1.32
		FA	-2.72	0.73	-0.32	0.00	1.32
	excluded	Age			(-0.10)	0.26	1.00
		NA			(0.06)	0.50	1.01
		FI			(-0.09)	0.31	1.02
		FI*FA			(-0.10)	0.38	1.54
		FI*FIQ			(-0.04)	0.33	1.07
		FA*FIQ			(-0.23)	**0.65**	**9.30**

*Adj. R^2 = 0.10
**Adj. R^2 = 0.20