Legible, are you sure? An experimentation-based typographical design in Safety-critical context

Jean-Luc Vinot
Université de Toulouse – ENAC - IRIT
7 av. E. Belin, Toulouse, France
jean-luc.vinot@enac.fr

Sylvie Athènes
Université de Toulouse – UPS, France
PRISSMH EA4561
athenes@cena.fr

ABSTRACT
Designing Safety-critical interfaces entails proving the safety and operational usability of each component. Largely taken for granted in everyday interface design, the typographical component, through its legibility and aesthetics, weighs heavily on the ubiquitous reading task at the heart of most visualizations and interactions. In this paper, we present a research project whose goal is the creation of a new typeface to display textual information on future aircraft interfaces. After an initial task analysis leading to the definition of specific needs, requirements and design principles, the design constantly evolves from an iterative cycle of design and experimentation. We present three experiments (laboratory and cockpit) used mainly to validate initial choices and fine-tune font properties. Results confirm the importance of rigorously testing the typographical component as a part of text output evaluation in interactive systems.

Author Keywords
Design, evaluation, experimentation, typography, legibility, readability, Safety-critical, aircraft, cockpit.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Design, Experimentation, Human Factors.

INTRODUCTION
When designing safety-critical interactive system, one needs to evaluate them against the requirements of Safety, Usability, Reliability and capacity to Evolve (SURE) [14] that can be seen as convergent and always desirable properties. Evolution in the design of safety-critical HMI calls for a change in the way we conceive of the links between user needs, system constraints, requirements and validation.

As part of a research cooperation with aeronautical industry, we were asked to go over text information visualization for all screens (cockpit and cabin) of future aircraft programs. The goal was to create a set of specially adapted fonts. Additionally, we had to establish the rules of use for these fonts geared to future cockpit interfaces, in order to provide design engineers with relevant information about the relationship between font specificities and context of use.

Cockpit screens, as others Safety-critical interfaces, are essentially based on the display of textual information and, thus, rest on the use of digital fonts. The large number of available digital fonts, as well as the published guidelines should not lead us to consider that legibility is no longer an issue of concern. On the contrary, a special effort should be made to prove the safety, usability and performance of this software component. The creation of a numeric typeface necessarily involves highly specialized knowledge in the field of design and typography. The critical area of use of these fonts also requires the contribution of particularly rigorous evaluation methodologies of the kind used by experimental sciences. In aeronautical context, design, development and operational deployment are strictly supervised by system engineering methodologies, evaluation and finally, technical and users' validation required for certification and approval of new operational systems.

This paper presents a study involving the design of typeface suited for cockpit, its development and evaluation. We present the different phases making up the study: task analysis, description of the specific needs, links with theoretical work, definition of the ensuing requirements and design principles and, lastly, an iterative process of design-experiment cycles geared to help and validate design choices. Among tests designed at the character, word or whole page levels, only results from three detailed experiments pertaining to character discrimination are reported here as a coherent whole within our design process. Finally, we conclude with a discussion of the contributions of this kind of integrated study within the design process and the possible implications for HCI.

PRELIMINARY ANALYSIS
This project required significant preliminary analysis: a phase of field observations of users in an operational context with pilots (simulator and real commercial flights), an expert analysis of the existing technical constraints, and a theoretical approach based on established knowledge

CHI 2012, May 5-10, 2012, Austin, Texas, USA.

from two domains, the perceptual processing of visual information and the typography.

Analysis of pilots activity

On commercial flights, pilot crews ensure five major activities: aircraft piloting (manually or with autopilot), navigation (managing and tracking flight route plan), aircraft systems monitoring, communication with air-traffic controllers and ground support, and lastly, management of the airline mission. In order to conduct these activities, they interact with specialized interfaces displayed on the cockpit screens. The increasing complexity of aircraft systems leads today to a more integrative approach with an aggregation of cockpit systems, and a significant increase in data and functions displayed for pilots. These complex visualizations may cause new risks to the usability of systems.

Observations of pilots in the cockpit have shown that both display environment and data have very special characteristics, thus entailing highly idiosyncratic needs. Below are listed the most important ones.

[**N1**] **Extreme lighting conditions**: pilots ceaselessly look from one screen to another in order to gather and compile information. Additionally, the visual transition between the aircraft inside and outside can be very abrupt. These constant adaptations to brightness and focus generate visual discomfort and fatigue.

[**N2**] **Viewing angles**: Multi-screen display in the cockpit implies that pilots have to read data viewed from different angles and distances.

[**N3**] **Time pressure**: Visual fixation dwells, which require more time for denser and more complex information [1], will however be shortened in heavy workload context [20].

[**N4**] **Specific data**: Texts are very specific: mostly symbolic and non-textual information, numerical values, system IDs, aeronautical abbreviations, labels or abbreviated sentences for menu or instruction lines, and even mixture of letters and numbers.

[**N5**] **Text legibility**: Text density can be quite overwhelming. Character design itself has some flaws that tend to bring about confusion between some letters. Significant problems were found concerning visual spacing, letter size or contrast. Lastly, fonts used on different views are sometimes heterogeneous. The close proximity or overlapping of graphical elements (highlights, framing, weather information, maps) can significantly interfere with legibility.

Results from this analysis allowed us to define a multidisciplinary process to design, guided by theory and validated progressively through experimental studies. We first researched relevant models from theoretical work in order to translate the previously defined needs into design requirements.

Related theoretical work

For anyone whose work implies presenting textual information, whether printed on paper or displayed on screen, there is a large corpus of knowledge and rules meant to facilitate text legibility, ranging from font size, colour and contrast choices, to spacing and text disposition. While making good use of this literature, the specific operational environment – aircraft safety-critical interfaces – drove us to explore more fundamental aspects of the reading mechanisms and use typographical knowledge and approach. On the one hand, time pressure, fatigue, extreme lighting contexts and heterogeneous complex data are seriously constraining text display. On the other hand, there is a much larger than usual expectation from the reader about the kind and whereabouts of the displayed information. Opening our study to new solutions meant fitting our specific reading context within the available knowledge about basic reading mechanisms.

During reading, the eyes move across the text, mingling short rapid movements and, about four to five times per second, short stops. This succession of saccades and fixations allows for groups of characters to be successively projected onto the central part of the retina, where they enter the nervous system to be further processed. This part of the retina, the fovea, is extremely small -1 or 2 degrees of visual angle- and is responsible for the most precise visual perception. In normal reading conditions, this area perceives at the same time 4 to 5 characters with great precision. Swift movements across the text, the saccades, tend to jump over 7/8 characters – the range being 1 to 20 characters – and serve to bring the fovea to the next relevant group of characters for a fixation. Saccade lengths, fixation durations and regressions to text already read vary considerably with reader experience and text difficulty.

The retina is not homogeneous, leading to considerable differences in the perceptual capacity of the central versus peripheral parts of the receptor field. We perceive precisely only the small part of the visual field which is projected on the fovea. Starting from this area, the rest of the visual field becomes progressively blurred [18][13]. Nevertheless, as far as reading is concerned, character size is not a decisive factor for the reading performance: smaller characters are not more difficult to read than bigger characters. Indeed, increasing the character size increases by the same token the peripheral area covered by any given number of characters, causing those falling outside of the fovea to become blurred. In contrast, smaller characters allow the eyes to perceive precisely a larger number of letters. Within limits of the retina resolution, the two mechanisms, central precision and blurred periphery, compensate each other, so that small and large characters tend to yield equivalent reading performances.

Saccade lengths tend to be rather constant in number of characters whereas they vary greatly in size, depending on the text font. Based on the number of characters –typically 7 to 9 – the brain anticipates the amount of eye movement

necessary for the saccade to jump over the right number of characters, regardless of their size. Seven to 9 seems therefore to be the amount of characters processed during a fixation. Processing time can be extremely brief. Although the average fixation duration is about 200ms, a display duration of about 50ms allows for reading to proceed normally [10]. While the fovea identifies exactly the characters, the peripheral vision uses their global shape in order to anticipate the words yet to be processed [12]. Obviously, some characters are easier for the peripheral vision to discriminate, for example, "i" and "p", or "l" and "m". Studies have proposed to cluster letters according to specific parts of their anatomy (*stem*, *ascender*, *descender*…) [6]. Most probably, when character shapes are very close (for example, "a" and "e"), the brain additionally uses linguistic regularities and semantic context to help anticipation. In fact, Reicher [16] showed a Word Superiority Effect where letters embedded within words are identified faster than when they are presented in isolation. More recent work by Pelli *et al.* [15] nuances this effect by showing that a word is never detected as a single feature but more likely as a set of simple features detected at the letter level. Such results support the idea that reading involves detection and likelihood of features at several levels at the same time, characters, graphemes, syllables and words, each level providing a linguistic context to help recognition. However, any Word Superiority Effect is of limited use for deciphering some of the specific data displayed in aeronautical context where legibility may rest crucially on the discrimination between individual characters in order to disambiguate non words.

Visual perception and spatial frequencies

Visual perception can be described in terms of spatial frequencies or, in other words, as a number of cycles per degree of visual angle. One can very simply visualize this notion as a grid of very thin black vertical lines on a light background: our perception is best when there are 8 lines (or cycles) per degree of visual angle. Above 50-60 cycles per degree of visual angle, most people perceive only a uniform grey. While being a good approximation of a standard situation, these results are not absolute and depend on display contrast. Any letter of a given size at a given distance from the eyes can be described in terms of spatial frequencies: the global shape will be defined by lower frequencies whereas fine details will come from higher frequencies.

Enhancing legibility

We already stated that the specific organization of the retina entails a decrease in visual precision from the fovea to the periphery. Another way to describe this phenomenon is to express this decrease in terms of a differing sensitivity to spatial frequencies: the fovea is sensitive to high frequencies whereas the periphery is sensitive to low frequencies. Spatial frequency is thus a very useful notion to describe and compare character shapes [17]. Basic character strokes are vertical, horizontal or diagonal, straight or curved. In a normal comfortable reading situation, the frequencies making up the character basic strokes are low, typically 6 to 8 cycles per degree of visual angle, while details, such as serifs, are high frequency components [17]. More recently, Majaj *et al.* [9] proposed the concept of stroke frequency, i.e. the number of lines crossed by a slice through a letter, divided by the letter width. As a result, global –lower frequency- shape of the characters will be well perceived by the peripheral retina, but details will be perceived only by the central retina (aka fovea). In order to help character discrimination by the peripheral retina, and therefore facilitate word anticipation and presumably reading, one has to pay attention to the low frequency components of the characters.

Enhancing readability

Highly readable displays allow for good anticipation and less demand on attention [20]. When the characters projected on the peripheral retina are too blurred to be read, the brain still uses the perception to detect alternating space and characters, and word length. In order to enhance further readability, one should pay attention to the relationship between the inner spaces of the character, the spacing between characters and between words.

Typographical means

The field of Typography produced a large corpus of knowledge and rules, commonly used by designers to address typographical needs of interfaces. Today, management and rendering technologies for digital text are very efficient. The available type library (digital fonts) for many languages and uses is large, but primarily intended for displaying text written as sentences. For our own needs, the display of specific textual information in Safety-critical context, we had to go back to the foundations of typography, letter anatomy, composition and harmony. We conducted a detailed anatomical study of alphanumeric characters. We described the structural elements of stroke and listed the main characteristics to be safeguarded for each glyph.

Letter Anatomy and typographic contrast

A font is a typographical representation of writing, consisting of a set of characters. In digital font, characters are instantiated by glyphs, images (graphemes) of typographical sign [4]. The shape of these typographical signs is built from regular stroke parts corresponding to basic gestures of the character writing. Five standard parts describe the academic drawing of most Latin letters anatomy: the *stem*, e.g. the vertical line of I character, the *arm* or *crossbar*, e.g. the horizontal lines of E or A, the *stroke*, e.g. the diagonal branches of A or K, the *bowl*, *stress* or *spine*, e.g. the curved strokes of O or S, the *leg* or *tail* of R or Q. Along with these five parts, there are many other anatomical elements such as *spikes* (A) or *loop* (g), *ascenders* and *descenders* (projecting parts of lowercase letters), or *serifs* (small endings, originally induced by the tool used for drawing a letter).

Typographic contrast is the variation in thickness between the thicker and thinner parts of the character. Some lineals (*sans serif*) typefaces, such as Arial, have a very low contrast, thus inducing simpler, more basic shapes, than *Oldstyle* type, like the Times New roman used for this text. Typographic contrast is used to reproduce writing gestures and is thus better suited to visually express the stroke parts making up a character.

Harmony, visual alignments and spacing

For typography, harmony is based on balance and visual consistency of form and composition. Character drawing is based on three main visual horizontal alignments: *baseline*, *cap-height* and *x-height* (lowercase). Traditionally, stroke and contrast regularities are produced by the writing implement, oriented to form a slanted axis for character drawing. For Typographers, spacing is one of the most important issues to address. The visual balance of inner spaces of each character shape (counter-forms) and the regular spacing between letters and words are fundamental. Due to the constraints of pixels at low resolution, counter spaces are especially difficult to adjust in digital fonts.

Respect for the letter anatomy and visual alignments, good use of typographic contrast and balanced spacing are essential to the legibility and harmony of typefaces [7].

Typographical invariants: the font properties

Digital fonts are categorized by properties, such as weight (regular, bold) or slant (roman, italic) called typographical invariants. A typeface includes a set of fonts that express these properties. Creation of fonts requires fine-tuning of these properties.

Legibility of digital fonts

Adrian Frutiger [3] worked for the signage system of Roissy-Charles de Gaulle Airport in France. He designed the typeface to ensure maximum text legibility, both from afar and from various angles. Among others, he tested the visual robustness of blurred characters. The high quality of Roissy typeface is due both to extensive work on the design of letter (large aperture, subtle optical corrections) and to a carefully balanced and aesthetical typography. This work led to the creation of the Frutiger typeface which was used for roadway signage in France and Switzerland and many transit systems around the world.

Microsoft's Verdana typeface [11], was designed by Matthew Carter and hinted by Tom Rickner specifically to address good readability of text on screen and rendering of scalable characters, even for low resolution devices. In these fonts, incorporating many hand-hinting instructions enhances pixel rasterization. They attempt to correct undesirable rasterization effects of glyph by equalizing the weights of *stems*, *arms* or *stroke* letter anatomy, thus preventing parts of glyphs from disappearing. As a consequence, even for low resolution, hinting maintains legibility and aesthetic appearance. More than the important work on shape and discrimination of characters, one of the unique qualities of Verdana is the regularity of the spacing (inside and between letters), producing an excellent readability.

French Air Navigation Services have conducted studies on the specific typographical needs for air traffic control interfaces in safety-critical context. Similarly to cockpit screens, the data displayed on the controller's radar screen are very specific. Especially tricky are callsigns (aircraft ID), a mixture of letters and numbers where two callsigns may differ by only one character, for example, AF974ZL and AF9747L. Both bitmap (ODS) and vector (Bleriot) digital fonts families have been designed to enhance legibility and ensure discrimination between characters. These fonts have been evaluated and recommended in a Eurocontrol study [5] and are currently used for ATC and aircraft systems.

FROM REQUIREMENTS TO DESIGN PRINCIPLES

Building on the needs defined in the activity analysis phase and the theoretical approach, we were able to list the requirements necessary to guide the design and, whenever possible, to highlight and scientifically express important issues. We then proceeded to design a first prototype of a cockpit suited typeface. These requirements relate to major aspects of typography traditionally defined by the terms of legibility and readability. In addition, we had to take into account more technical requirements to comply with current layout constraints of cockpit screens.

Requirements

[**R1**] **Legibility** concerns more specifically the form and the visual rendering of each individual character. Particularly in Safety-critical systems and for specific data, each glyph, even isolated, must be completely identifiable and easily discriminated from other glyphs. As already stated, glyphs should ideally be distinguished from one another solely on the basis of low-frequency visual information [17].

[**R2**] **Reading performance** must be maintained in spite of short visual fixation durations and be resistant to low degrees of visual angle and high angular distortions.

[**R3**] **Robustness** in degraded visual environment is necessary to address the risks of high external light changes leading to a significant loss of text contrast. In complex graphic context, it is also necessary to address multi-layout interfaces and text superposition with other graphical objects.

[**R4**] **Readability** results from the complete process of presenting textual material in order to communicate meaning. In typography, readability strives to improve reading efficiency through coherence and regularity of text disposition, and letter proportions and contrast.

[**R5**] **Reading comfort** concerns the regularity of shapes, visual alignments, spacing and composition of characters in text string [10], and should allow for better planning of visual saccades.

[**R6**] **Compliance** with the layout constraints of current cockpit display models refers to the capacity to maintain,

for example, text sizes and characters per line density, strict vertical alignment for each character of a given displayed data, or specific charset.

[**R7**] **Semantic soundness** entails that the typeface, in a general sense, must be in harmony with the meaning of aeronautics.

Translating these requirements into typographic design principles is difficult because these set of requirements are partly conflicting or highly dependent. Thus, comfort of reading requires regular shapes whereas legibility in aeronautical context requires increasing visual distance between glyphs; or compliance with current text density on cockpit screen coerces character width and can interfere with the legibility or readability. Of all the requirements listed above, only R1 to R6 have been tested ; R7 has been taken into account in the design but not tested.

Design principles

We have identified a set of typographical recommendations for design, not strictly limited to, but including:

- Typeface will strictly express basic stroke parts of letter anatomy (in accordance with our anatomical study) to provide *good character identification* (R1, R2, R3).
- The shape of each alphanumeric letter will be particularly differentiated from the shape of other characters with which it could be confused (R1).
- Links and junctions between basic stroke parts of each letter will be carefully drawn to ensure clarity at low resolution and increase *reading performance* (R2).
- Numeric characters will use specific forms such as slashed "0", open form of "4" or a large hook for "1", to ensure that numbers will be perceived as a separate set and *not be confused* with capital letters (R3).
- Character width will be carefully reduced as condensed font form in order to *ensure text density compliance* (R6) while being visually *robust to angular distortion* (R2).
- Width of numeric characters will be strictly equalized to allow *vertical alignment* of numeric values (R6).
- Typographical contrast (thickness variation of character strokes) will be strong enough to guarantee symbol unity and *robustness in complex graphic context* (R3).
- X-height alignment of the font will be low enough to yield a good contrast of lowercase ascender and to improve *text prediction in peripheral vision* (R1, R2).
- Kerning will be specially adjusted to display short words and alphanumeric values in *compliance with aeronautical needs* (R4, R6).
- Fonts will use hinting instructions to ensure *good display for low-medium resolutions* (R2, R5, R6).
- Typeface will be both visually stern and aesthetic to satisfy pilots and be in *harmony with the semantic of aeronautics* (R5, R7).

FIRST DESIGN PHASE

We used the above listed requirements and design principles to conduct a first typographic design and produce a sofware prototype of the font.

Analysing readable fonts

We first performed a morphological comparison of sans serif fonts reputedly designed to maximize readability in order to analyze their forms and properties.

Figure 1: superposition of alphanumeric glyphs from 10 readable fonts, right, glyphs "Q" and "3" details.

Figure 1 was produced by superimposing the glyphs of ten digital fonts: Univers, Frutiger, Helvetica Neue, Vera Sans, Verdana, Lucida Grande, Myriad Pro, Calibri, Tiresias PC and Bleriot. Overlapping of glyphs shows a high overall similarity of forms, with some interesting variations for some signs, for example glyphs "Q" or "3" (details on Figure 1). Proximity of the main typographical invariants values (weight, character width) of these fonts allows us to consider them as reference values for our experimental exploration of font properties.

Regardless of their qualities, none of these fonts satisfies all the requirements. The most difficult requirement is compliance with cockpit layout constraints (R6).

Exploring solutions

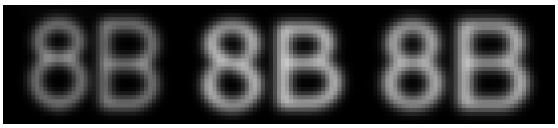

Figure 2: visual comparison after applying a Gaussian blur on the glyphs "8" and "B" of three fonts.

One of our design principles is to discriminate glyphs on the basis of low visual frequencies and thus improve the prediction of the characters in peripheral vision. For a low-resolution text rendering, rasterization of characters adds blur to the glyphs (antialiasing) and smooth angles of visual forms. Figure 2 illustrates this visual effect by applying a Gaussian blur on a pair of characters (8, B), displayed using 3 different fonts: aeronautical, Verdana and Helvetica Neue.

Figure 3: half-bitting rendering technique (left); two modified glyphs blurred and not blurred (right)

Rubinstein [17] proposes the use of the *Half-Bitting* technique to improve the rendering of printed text (laser

printer) with more subtle visual strokes. This technique allows, for example, to increase an angle by adding or removing pixels at opposed outgoing and incoming angles of the form (figure 3, left), in a fashion similar to the one used by artists to increase the salience of angles by extending or distorting the edges. The rendering visually enhances the characteristics of the expected form.

We have used special serif forms, like glyphic (*incised*) typefaces, to test the visual increase of angles and junctions of stroke parts. This form is compatible with the design of lineals sans serif fonts and can be achieved by the creation of light serifs, thickening the extremities of shapes like the *Optima* fonts. Figure 3 (right) shows two modifed glyphs of the first prototype (with and without Gaussian blur). Even after blurring, the glyphs remain visually well discriminated.

Designing the typeface

Taking into account the previous remarks and requirements, the next step was to instantiate a typeface. This typographic creation rests on a calligraphy work, calling for use of a drawing tool on paper. After a paper study with roughs, a reduce set of characters was drawn in large format, then scanned and vectorized. Bezier paths of letters were finalized and harmonized. Using these vector shapes, a first TrueType font (figure 4) was then created with the outline font editor FontForge, including uppercase and lowercase alphabets, numbers, and ASCII punctuation and symbols.

Figure 4: first font prototype (numbers and capitals letters)

ITERATIVE CYCLES OF DESIGN AND EVALUATION

Once this first prototype achieved, we developed a continuous interactive process of design and experimentation in order to explore potential typographic solutions, to verify the requirements and to progressively adjust the graphic properties of typeface.

Experiment 1

The goal of the first experiment was to position our prototype in terms of legibility and character discrimination. We thus compared a font widely used in aeronautical context, a font well-known for its good readability (Verdana) and our first font prototype. In order to enhance the role played by low-frequency components in character recognition, the stimuli were displayed with a low character/background contrast which was adapted for each subject with the help of a pre-test.

Twelve subjects (aged 21 to 47) named characters (letters and numbers) which were briefly presented on a screen (LCD 30" Apple Cinema HD display) set either 80cm or 100cm away from the subject. A device bolted to the table maintained the subject/screen distance.

Pre-test

Each session started with the above-mentioned pre-test: seated 100cm away from the screen, the subject was briefly presented with a Landolt C (a ring with a gap oriented in various positions) and had to report its orientation (top, bottom, left or right) using the arrows on the keyboard. The size of this symbol was 4.75mm, the same as the character size used in the experiment. Twenty contrast values were used for the test: starting with a good contrast, each following symbol was stepwise displayed in a decreasing contrast value. After the 20 decreasing steps, the process was reversed and, starting with the last, virtually unperceivable, contrast value, it was stepwise increased back to good contrast values. A sound was emitted when the symbol was displayed and if the subject could not perceive its orientation, s/he was to report it by depressing the space bar. Results were scanned and an appropriate contrast value was chosen in the response range between "always correct" and "never perceived". This contrast value was used throughout the subsequent experiment.

Experimental task and settings

The subject had to name as quickly as possible the character which was displayed on the screen center. "Not identified" was also a possible response. For each font (aeronautical, Verdana and the prototype), thirty-six characters (26 letters: A to Z; 10 numbers: 0 to 9) were presented. Figure 5 illustrates the three fonts. A trial started with the display of a fixation pattern on the screen center during 700ms, followed by the display of the character at the same location during 17ms, then followed by a 200ms-span without display, during which time the subject gave the response. The next trial started with the display of the fixation pattern. A block consisted of the successive display of the 36 characters for one given font. Within each block, the characters were pseudo-randomly presented. For each experimental condition, there were 2 blocks. The subject could rest at will at the end of any given block.

Independent variables

There were 2 independent variables: the 3 fonts (aeronautical, Verdana and prototype) and the 2 subject/screen distances (standard 80cm and more difficult 100cm). The font and distance variables were counterbalanced on the subjects.

Figure 5: glyphs E, 0 and 1 from three tested fonts (from left to right, aeronautical, Verdana and prototype)

Dependent variable

There was only one dependent variable: the subject's response which could be "correct", "wrong" or "no-response" (meaning that the subject could not identify the character).

Results

Results are presented on Figure 6. For each type of responses (correct, wrong and no-response), an ANOVA with a repeated measures within subject design (3 fonts x 2

distances) was performed. The ANOVAs showed significant effects of the distance: the larger distance (d2) entailed an increase of the number of no responses ($F(1,11) = 27.16$, $p = 0.000$), an increase of the number of wrong responses ($F(1,11) = 62.14$, $p = 0.000$) and a decrease of correct responses ($F(1,11) = 125.37$, $p = 0.000$). Sustained by significant pairwise comparison, the ANOVAs showed also an effect of the font: Verdana and the prototype font both produced less no responses than the aeronautical font ($F(2,22) = 35.52$, $p = 0.000$); the prototype font gave rise to more correct responses ($F(2,22) = 104.79$, $p = 0.000$) and less wrong responses ($F(2,22) = 25.39$, $p = 0.000$) than Verdana, which itself fared significantly better than the aeronautical font. Significant interactions showed that the effect of the distance was greater for the aeronautical font, decreasing the number of correct responses ($F(2,22) = 4.3$, $p = 0.042$) and increasing the number of no responses ($F(2,22) = 24.44$, $p = 0.000$). In contrast, for the prototype font and Verdana, the effect of the distance was larger on the number of wrong responses ($F(2,22) = 4.24$, $p = 0.028$).

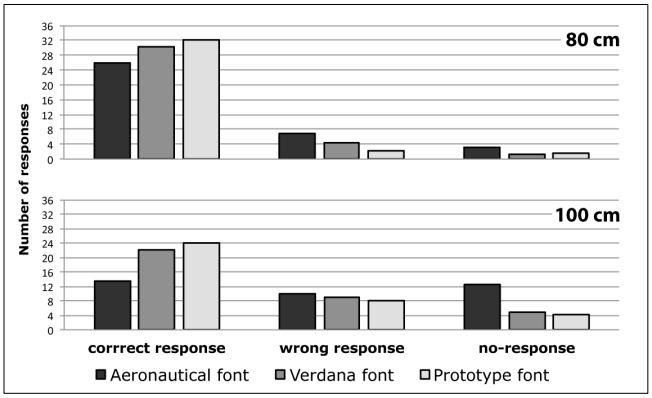

Figure 6: Experimentation 1, distribution of responses (averaged over all subjects) for subject/screen distance of 80 cm and 100 cm.

In other words, with increasing distance, the aeronautical font became illegible whereas the prototype font and Verdana gave rise to more errors. To sum up, the prototype font and Verdana produced, by and large, results which were fairly similar and significantly better than the aeronautical font. Though of a different magnitude, there was also a slight, but significant advantage for the prototype font over Verdana.

Confusion matrix

In order to better understand the role of features for letter recognition, we generated letter confusion matrices [2]. Such matrices show the relationship between the displayed stimulus and the response given. It is a somewhat indirect but useful measure of character shape similarity. Figure 7 shows detail of a matrix.

By and large, for the aeronautical font, the distribution of confusion between characters is quite spread out, many confusions being made only once; no-responses are numerous and correct responses range from 2 to 14. These results reflect the great difficulty in discriminating the glyphs, especially at the larger distance. In contrast, Verdana and the prototype font confusion matrices yield tighter distributions, reflecting mostly known confusions between characters. For prototype font, major confusions are: I for 1 (9 occurrences), S for 5 (6), Z for 7 (6) or Z for 2 (4), N for H (5).

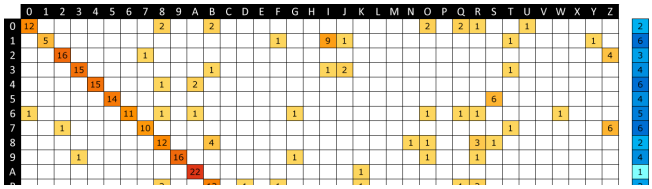

Figure 7: letter confusion matrix (detail) for the prototype (100cm). Left column shows displayed characters; top row shows subjects' responses. Each intersection of column and row presents the number of occurrence of the stimulus/response pair. The correct responses are displayed on the diagonal of the matrix (stimulus "A" and response "A"...). All other cases are wrong answers, e.g. stimulus "5" and response "S". The far right column (on blue) presents the number of no-responses for each character.

Experimentation 2

The goal of this second experiment was to repeat our first experiment within the context of a cockpit, using the displays, orientations and distances from the screens as they are in aircrafts. We thus compared the same 3 fonts as in experiment 1 (a font widely used in aeronautical context, Verdana and our prototype), using the same experimental design. In the following section, the experimental information will be reported in detail only when departing from experiment 1.

Twelve subjects (aged 25 to 35) named characters (letters and numbers), which were briefly presented on a screen. As in the first experiment, the stimuli were displayed with a low character/background contrast, which was adapted for each subject with the help of a pre-test.

Experimental task and settings

The only departure from the settings of experiment 1 was the length of time during which the characters were displayed on the screen. Due to technical constraints, in order to ensure display of the character on cockpit screens, it had to be displayed over 2 cycles, thus during 34ms (as opposed to 17ms for experiment 1). Independent variables and dependent were the same as in experiment 1.

Results

Results are presented on Figure 8. For the "correct" and "wrong" type of responses, an ANOVA with a repeated measures within subject design (3 fonts × 2 distances) was performed. The number of "no-response" was to low to allow for an ANOVA, so that when reasonable a chi2 was calculated. The ANOVAs showed significant effects of the distance: the larger distance entailed an increase of the number of wrong responses ($F(1,11) = 51.34$, $p = 0.000$) and a decrease of correct responses ($F(1,11) = 129.26$, $p = 0.000$). The ANOVAs showed also an effect of the font: the prototype font gave rise to more correct responses ($F(2,22)$

= 64.04, p = 0.000) than both Verdana and the aeronautical font (post hoc pairwise comparisons show significant effects, p = 0.000) and less wrong responses (F(2,22) = 30.19, p = 0.000) than Verdana, which itself fared significantly better than the aeronautical font. Results from the chi2 on the "no-responses" showed a significant effect of the font (Chi2(2) = 24.62, p < 0.000). Significant interactions showed that the effect of the distance was greater for the aeronautical font, greatly decreasing the number of correct responses (F(2,22) = 11.52, p = 0.000) and increasing the number of wrong responses (F(2,22) = 3.59, p = 0.045), as well as the number of no-responses (Chi2(5) = 185.47, p < 0.000).

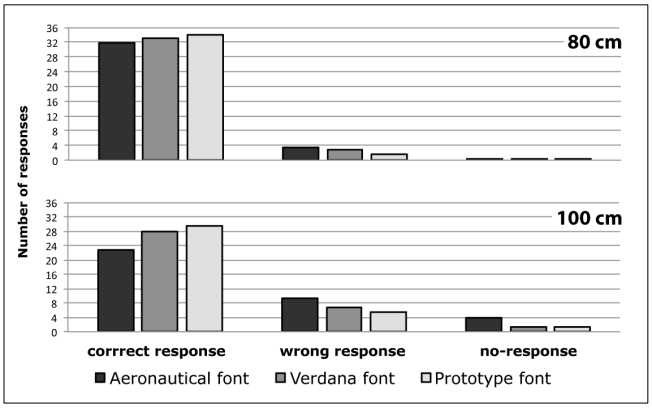

Figure 8: Experimentation 2, distribution of responses (averaged over all subjects) for subject/screen distance of 80 cm and 100 cm.

Comparison between experiments 1 and 2:

We found the same trends in both experiments where the aeronautical font scores notably worse than both Verdana and the prototype. The main difference between the two experiments lies in the overall level of performance. Due most probably to the longer display time (34ms instead of 17ms), results from the cockpit simulation show a larger amount of correct responses and a smaller amount of wrong or no-responses. It is interesting to note that, whereas the decrease of no-responses in the cockpit setting is of the same magnitude for all 3 fonts (78% on average), the increase of correct responses, as well as the decrease of wrong responses set the aeronautical font apart. Indeed, the increase of correct responses is on average greater for this font (45% versus 14% and 17% for the prototype and Verdana respectively). The increase reaches 67% for the aeronautical font at the larger distance, versus 23% and 26% for the prototype and Verdana respectively. Concerning the decrease of wrong responses, it is on average the same for all 3 fonts (about 28%). However, whereas the decrease is of the same magnitude for the prototype and Verdana irrespective of the distance (between 26% and 31%), it varies from 48% at the smaller distance to 5% at the larger distance for the aeronautical font. In other words, the longer display time of the stimuli in the cockpit helped all fonts, but it did so to a greater extent for the aeronautical font. Furthermore, for this last font, the larger distance still proved to be an hindrance that the longer display time did not compensate. Noteworthy is the fact that the rough prototype fares slightly but consistently better than Verdana.

Experimentation 3

The goal of this experiment was to help the design in setting a correct value for the character weight, taking into account that the characters would be displayed using different polarities (on black, white or gray backgrounds). In order to evaluate the robustness of each tested weight, we used 2 character/background contrast values. In this experiment, we used only instances of the prototype font.

Twelve subjects (aged 21 to 57) performed a visual search task (letters and numbers). At the end of the session, using the pairwise comparison method, subjects were shown samples of the characters from the experiment and asked for their preferences.

Experimental task and settings:

The subject was presented with a 10 x 10 characters table displayed on a screen (LCD 30" Apple Cinema HD display) placed 80cm away and had to search for a given target character among distractor characters. Once the search was completed, the subject gave the number of occurrences he thought was correct. There was no limit to the search time, even though the instructions emphasized speed and accuracy. A trial started with the display of a round target pattern on the screen center during 500ms, followed by the display of the character to be searched during 1s. The 10 x 10 table was then presented centered on the screen. When the subject was finished, s/he depressed the space bar, which caused the table to disappear from the screen, and announced the number of occurrences s/he had found. For any given trial, the number of occurrences of the target character could pseudo-randomly be 1 to 5.

Taken from our prototype font, there were 36 distractor characters (numbers: 0 to 9; letters: A to Z) and 18 target characters, numbers and uppercase letters (0, 1, 2, 4, 6, A, B, E, G, H, I, K, M, N, O, R, S, Z), chosen for their tendency to become blotched with increasing weight (for example, E or A), or their capacity to be confused with other characters (for example, 1 and I, or O and 0). Thus, any given table was made of 1, 2, 3, 4 or 5 occurrences of a given target character and the completing number of distractor characters (99 to 95). Each target character could pseudo-randomly appear in any given contrast/weight condition, but was presented once and only once in each polarity.

Independent variables

There were 4 independent variables: 3 polarities (white, black or gray backgrounds), 2 contrast values (normal or low), 3 weights (heavy, normal or light) and 3 repetitions. Polarity, contrast value and character weight were blocked but counterbalanced on the subjects. The 3 repetitions of each experimental condition were blocked.

Dependent variable

There were 2 dependent variables, the subject's response time and the difference between the number of occurrences of the target character and the reported number thereof.

Results

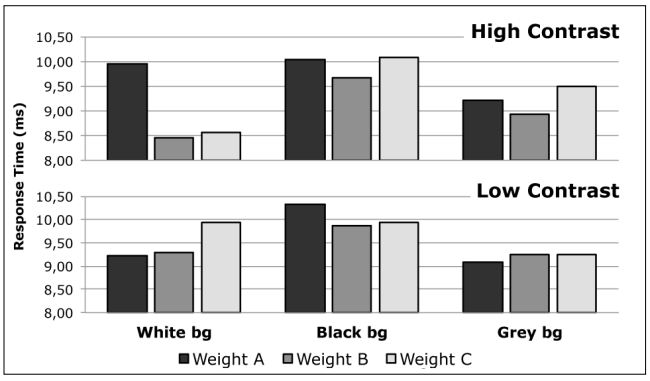

Figure 9: Experimentation 2, average response time for each font weight on 3 backgrounds with high or low contrasts.

Results are presented on Figure 9. The ANOVA on the response time showed only 2 significant main effects and no significant interaction effect. There was an effect of the polarity ($F(2,22) = 3.778$, $p = 0.038$) where post-hoc pairwise comparison showed that response time was longer for the black background than for the white background. The gray background was not significantly different from, either the white, or the black background. The significant effect of the contrast ($F(1,11) = 4.968$, $p = 0.047$) showed that response time was longer when the contrast was lower. Lack of further significant results is probably due to the fact that the range of the weight values was purposefully chosen very small. Indeed, larger differences might have yielded significant effects but were of no interest within the context of our study. One should keep in mind that the goal of this experiment was to help the designer choose the correct value(s), not to prove the effect of a given factor.

Pairwise comparison

At the end of the session, we displayed on the screen pairs of pseudo-words or numbers, on the same background, but using different weights. For each pair, subjects were asked to tell which seemed more pleasing and comfortable to read [19]. Results show that heavier and medium weights were preferred on black and grey backgrounds whereas medium and light weights were preferred on white background. In other words, if choosing only one weight value, the medium one is best to make reading comfortable regardless of the polarity.

DISCUSSION

In this study undertaken in the context of safety-critical systems and difficult environments, we integrated design, scientific models and experimental observations in order to elaborate typeface artifacts in a carefully controlled stepwise fashion. Mackay & Fayard [8] provided a framework for interfaces design, integrating research, engineering and design. They describe how interaction models can be created from theory and observations to instantiate new artifacts, ranging from early simulations to working prototypes to products. The design team must agree to work in a very interactive but constrained manner, each domain imposing or adapting its own rules and limits. In our study, design and typeface artifact constantly evolve from experimentations. To conceive the experiments, we had to operationalize the requirements of legibility with typographic properties and experimental questions. For example, testing robustness entailed manipulating the weight properties to ask if more pixels lead to better shape perception, especially in degraded environment (low contrast). Thus, needs, requirements and design choices are linked through quantified typeface parameters (for example, typographic contrast, weight, spacing…). With this method, designing takes time but yields a multi-layered product which can easily be further adapted, should the need arise.

Safety-critical context of use has driven our choice to go through iterations of short cycles of experimentation-based design in order to ensure conformity with previously defined requirements and design principles. These experiments were designed either to validate broad typographical choices or to fine-tune font properties.

Validating Design

Results from experiments 1 and 2, a comparison of font legibility, has shown that our basic prototype font does well with respect to the tested requirements of legibility (character recognition and discrimination) and robustness, and yielded better results than an in-use aeronautical font and the Verdana font. We conducted these experimental manipulations in order to validate our initial design concepts, first in a carefully controlled laboratory environment and, then, in a real but controlled cockpit environment. To our knowledge, the replication of laboratory experiment in an operational environment, as well as the transposition of laboratory experimental constraints to an "almost" real world cockpit setting is fairly innovative. It allowed us to quantitatively evaluate legibility and validate our results on the intended destination screen and in cockpit ergonomic context (pilot position, viewing angles and lighting). Indeed, this type of experiment should be replicated on the final product if one wants to definitely state its legibility as required.

Helping Design

Even though our font design was largely validated, the confusion matrices pointed to legibility problems for some character shapes. Confusion matrices reveal not only the localization but also the direction of asymmetric confusions. For example, in the 5 and S pair confusion, we found that the problem was rather with the drawing of "5" than that of the "S". Used on results from experiments geared towards design validation, confusion matrices nevertheless helped refine character shape.

Experiment 3 is but an example of the experiments currently carried out on the typographic properties of the prototype font, such as weight, slant, and spacing among

others. Comparing instances of the prototype font designed with different values of one given property, has allowed us to build a range within which the effect of the values is known. Not only does it help to choose the best values for a given effect, but it also allows predicting with fair confidence the consequences on a given property to have to contend –for some reasons- with "less than best values". In other words, such fine-tuning of font properties serves to define a space of available choices, and their effects, for the design. In contrast with traditional end-user evaluations which strive to validate a finished product with respect to a definite set of specificities, our iterative evaluation process leaves room and direction for changes, should the requirements change. We can also use this range of values to ensure compliance across multiple requirements. For example, the manipulated range of available width helped design condensed shapes to address the text density of current cockpit interfaces.

Implications for HCI

The experimentation-based design process was necessitated by the safety-critical context of use. In mundane context, digital typefaces tend to be taken for granted. Reading is such an ubiquitous task that it is rarely tested as such. Generally, text reading is the unquestioned input for a given tested interaction involving a widget, for example. The ensuing observed performance will be understood as a function of the widget. As long as text display and rendering conforms to set specifications, interfaces designers will probably not subject the chosen typefaces to the rigors of testing and evaluation they subject other aspects of interface design, let alone tinker with the typefaces. Reading is a task requiring cognitive resources and typeface legibility is bound to influence cognitive load. Our results clearly show how minute design differences (such as half-bitting rendering technique) can influence character reading performance, even when comparing two typefaces, both aiming above all for legibility, such as Verdana and our prototype. Acknowledging the importance of typefaces and validating their usability should not be confined to safety-critical contexts. The ubiquity of text display in numerous interfaces and the multiplicity of contexts lead us to believe that the experimental verification of such typographical components in context of use is fundamental for interfaces design.

REFERENCES

1. Bellenkes, A.H., Wickens, C.D. & Kramer, A.F. Visual scanning and pilot expertise: The role of attentional flexibility and mental model development. *Aviation, Space, and Environmental Medicin, 68*: 569-579, 1997.
2. Bouma, H. Visual recognition of isolated lowercase letters. *Vision Research 11*: 459-474, 1971.
3. Frutiger, A. L'Homme et ses signes. Atelier Perrousseaux Editeur, 1999.
4. Haralambous, Y. Fonts & Encodings. Published by O'Reilly Media Inc, 2007.
5. Directed by Jackson, A. CoRe Project - Baseline Exemplary Style Guide, *Eurocontrol EATM Guidelines - HRS/HSP-006-GUI-01*, 2004.
6. Jacobs, A.M., Nazir, T.A. & Heller, D. Perception of lower letters in peripheral vision: A discrimination matrix based on saccade latencies. *Perception & Psychophysics 46*(1): 95-102, 1989.
7. Larson, K., Hazlett R.L., Chaparro, B.S. & Picard, R.W. Measuring the aesthetics of reading. *People & Computers XX – ENGAGE,* In *Proc. HCI 2006,* Springer (2006), 41-56.
8. Mackay, W.E. and Fayard, A-L. HCI, Natural Science and Design: A Framework for Triangulation Across Disciplines. In *Proc. ACM DIS '97,* Amsterdam, Pays-Bas: ACM/SIGCHI (1997), 223-234.
9. Majaj, N.J., Pelli, D.G., Kurshan, P. & Palomares, M. The role of spatial frequency channels in letter identification. *Vision Research 42*, 1165-1184, 2002.
10. McConkie, G.W. & Rayner, K. The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics 17*: 578-586. 1975.
11. www.microsoft.com/typography/web/fonts/verdana/
12. Nazir, T.A., Heller, D. & Sussmann, C. Letter visibility and word recognition:The optimal viewing position in printed words. *Perception & Psychophysics 52*(3): 315-328, 1992.
13. Nazir, T.A., O'Reagan, J.K. & Jacobs, A.M. On words and their letters. *Bulletin of Psychonomic Society 29*(2), 171-174, 1991.
14. Palanque, P., Basnyat, S., Bernhaupt, R., Boring, R., Johnson, C., and Johnson, P.Beyond, usability for safety critical systems: how to be sure (safe, usable, reliable, and evolvable)? In *CHI '07,* ACM Press (2007), 2133-2136.
15. Pelli, D.G., Farell, B. & Moore, D.C. The remarkable inefficiency of word recognition. *Nature 423*, 752-756, 2003.
16. Reicher, G.M.. Perceptual recognition as a function of meaningfulness of stimulus material.*Journal of Experimental Psychology,* 81, 274-280, 1969.
17. Rubinstein, R. Digital Typography: an Introduction to Type and Composition for Computer System Design. Addison-Wesley Longman Publishing Co., Inc. 1988.
18. Sere, B., Marendaz, C. & Herault, J. Nonhomogeneous resolution of images of natural scenes. *Perception 29*(12): 1403-1412, 2000.
19. Thurstone, L.L. A law of comparative judgement. *Psychological Review 34*: 278-286, 1927.
20. Wickens, C.D. & Hollands, J.G. Engineering Psychology and Human Performance. 3rd ed., Prentice Hall, New Jersey, 1999.