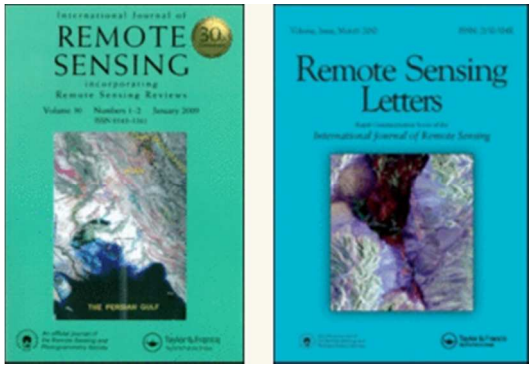


Variability of operator performance in remote sensing image interpretation: the importance of human and external factors

Journal:	*International Journal of Remote Sensing*
Manuscript ID:	TRES-PAP-2013-0603.R1
Manuscript Type:	IJRS Research Paper
Date Submitted by the Author:	18-Nov-2013
Complete List of Authors:	Van Coillie, Frieke; Gent University, Department of Forest and Water Management Gardin, Soetkin; Gent University, Department of Forest and Water Management Anseel, Frederik; Ghent University, Department of Personnel Management, Work and Organizational Psychology Duyck, Wouter; Ghent University, Department of Experimental Psychology Verbeke, Lieven; Gent University, Department of Information Technology De Wulf, Robert; Gent University, Department of Forest and Water Management
Keywords:	interpretation, image analysis
Keywords (user defined):	operator performance, image interpretation, human factors

Variability of operator performance in remote sensing image interpretation: the importance of human and external factors

FRIEKE M.B. VAN COILLIE*†, SOETKIN GARDIN†, FREDERIK ANSEEL§, WOUTER DUYCK‡, LIEVEN P.C. VERBEKE¶ and ROBERT R. DE WULF†

†Laboratory of Forest Management and Spatial Information Techniques, Ghent University, Coupure Links 653, 9000 Ghent, Belgium

§Department of Personnel Management, Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

‡Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

¶Department of Information Technology, IBCN, Ghent University, Gaston Crommenlaan 8 bus 201, 9050 Ghent, Belgium

*†Laboratory of Forest Management and Spatial Information Techniques, Ghent University, Coupure Links 653, 9000 Ghent, Belgium – T +32 9 2646108 – F +32 9 2646240 – E frieke.vancoillie@ugent.be

Operator performance in remote sensing image analysis: the impact of human and external factors

This study tackles a common, yet underrated problem in remote sensing image analysis: the fact that human interpretation is highly variable among different operators. Despite current technological advancements, human perception and interpretation are still vital components of the map-making process. Consequently, human errors can considerably bias both mapping and modelling results. In our study we present a web-based tool to quantify operator variability and to identify the human and external factors affecting this variability. Human operators were given a series of images and were asked to hand-digitize different point, line, and polygon objects. The quantification of performance variability was achieved using both thematic and positional accuracy measures. Subsequently, a series of questions related to demographics, experience, and personality were asked, and the answers were also quantified. Correlation and regression analysis was then used to explain the variability in operator performance. From our study we conclude that 1) humans were seldom perfect in visual interpretation, 2) some geographic objects were more complex to accurately digitize than others, 3) there was a high degree of variability among image interpreters when hand-digitizing the same objects, and 4) operator performance was mainly determined by demographic, non-cognitive and cognitive personality factors, whereas external and technical factors influenced operator performance to a lesser extent. Finally, the results also indicated a gradual decline in performance over time, mimicking classical mental fatigue effects.

Keywords: interpretation, image analysis, variability, human factors

1. Introduction

When examining centuries old maps, we tend to interpret them with caution. While we acknowledge that cartographers of the time were limited to only the use of their eyes and mind, we are often more surprised by the high quality of some of these maps than by the expected and unavoidable errors. Over the last decades, map-making technology has drastically improved and now includes advanced Remote Sensing (RS) and

Geographic Information Systems (GIS), opening a wide range of possibilities to diminish the influence of human errors. These technological advancements however provide the end-user with a false sense of security towards the quality of spatial information simply because human perception and interpretation still are vital components of the map-making process. Early in this process, for example, when images are registered, the operator is expected to accurately and precisely localize ground control points. Later in the process, maps are inferred from the images using algorithms that in turn require training data and parameter tuning, both of which are fairly subjective operator tasks. Finally, when it comes to algorithm validation and accuracy assessment of the produced maps, the human operator again intervenes. Typically, an accuracy evaluation of some automated process is performed by comparing the mapping result with ground or reference data prepared by an operator and assumed to be an accurate representation of reality. In fact, the ground data are just another classification which may contain both thematic errors and errors due to mislocation (Foody 2002). By considering operator-produced reference data to be perfect, with zero variance between judges, the human operator is implicitly assumed to be infallible. Foody (2002) questions this assumption by stating that an algorithm or classification accuracy assessment is actually only reflecting the degree of correspondence with these ground data, but not necessarily with reality. As any system is only as strong as its weakest link, human errors are still influencing the reliability of current advanced mapping technology.

In this paper, we first discuss the problem associated with human errors in the map-making process, followed by a review of the literature that has examined this problem. Section 2 of this paper presents the method we developed to address this problem and to achieve the postulated research objectives. In Section 3 we present and

discuss the results. Finally, we conclude by summarizing our major findings and propose a number of future research options.

1.1. Human image interpretation in remote sensing

Image interpretation is the act of examining photographic images for the purpose of identifying objects and judging their significance (Colwell 1997). Although most individuals have substantial experience in intuitively interpreting conventional photographs, the interpretation of aerial and space images departs from everyday image interpretation in three important aspects: (1) the portrayal of features from a downwards, often unfamiliar, perspective; (2) the frequent use of wavelengths outside of the visual portion of the spectrum; and (3) the depiction of the Earth's surface at unfamiliar scales and resolutions (Lillesand *et al*. 2008). Principles of image interpretation have been developed empirically for more than 150 years. The most basic of these principles are labelled as the elements of image interpretation. They include: shape, size, pattern, tone/colour, texture, shadow, location, height/depth and site/situation/association (Lillesand *et al.* 2008). These elements are routinely used when interpreting aerial photographs or space images. Well-trained image interpreters use many of these elements during their analysis without consciously considering them as separate items. However, novices may not only have to force themselves to consciously evaluate an unknown object with respect to these elements, but also to analyse its significance in relation to the other objects in the scene. Based on these basic elements, image interpretation keys are generally developed as training aids for novice interpreters or reference/refresher materials for more experienced users (Lillesand *et al.* 2008). Successful image interpretation is also coupled with personal characteristics of the interpreter. Lillesand *et al.* (2008) state that the most capable image interpreters

have keen powers of observation coupled with imagination and a great deal of patience. In addition, it is important that the interpreter has a thorough understanding of the phenomenon being observed as well as knowledge of the geographic region under study (Lillesand *et al*. 2008). They compare the image interpretation process to the work of a detective trying to put all the pieces of evidence together to solve a mystery. Depending on the level of complexity, the mystery-solving process varies from direct recognition of objects in the scene to inference of site conditions. Hence, the interpreter uses the process of *convergence of evidence* to successively increase the accuracy and detail of the interpretation (Lillesand *et al.* 2008). Nevertheless, every interpreter would use a different interpretation key to solve the mystery and not all interpreters would arrive at to the same conclusion.

The fact that different humans might interpret the same image in a totally differently way is the most common, yet at the same time, the most understated problem in human image interpretation (Albrecht 2010, Madden 2010). Within this view, Scepan (1999) initiated a new procedure in RS image analysis. For the land cover classification of high resolution image data, every image was handled by three different operators. At least two of the operators had to agree on a site before it could be assigned to a specific class. The utility of this approach was proven immediately by the number of sites over which there was disagreement. Another factor that was taken into account was confidence. When operators were asked how confident they were, they showed only medium confidence in their own work. This procedure has been used in other studies (Scepan *et al.* 1999; Zhu *et al.* 2000; Sarmento *et al.* 2009) where even lower confidence was shown (Zhu *et al.* 2000) or where the intervention of the third interpreter was necessary in 30% of the cases. If this had been a research on change detection, just the use of two different operators could have led to an observed change

of 30% of the land cover while in reality nothing had changed. Also, one can ponder the question whether an operator would come to the same conclusion interpreting the same image a second time. This raises only more questions about interpreter consistency. Similar procedures were also carried out in other studies. Leckie *et al.* (2003) used a procedure where two operators both interpreted the same images in order to define stand boundaries. In case of disagreement, discussion or even inspection of the automated boundaries was used in order to come to a consensus. The final decision was made after a third operator agreed upon the result.

Given that human image interpretation is still topical in the current map-making process, it is surprising that, in spite of its relevance, virtually no research has focused on operator functioning within remote sensing applications. This contrasts strongly with many other domains (see below) requiring similar human intervention, in which a sizeable amount of research has been carried out to investigate operator performance.

1.2. Psychological reasoning behind human image interpretation

The limited number of contemporary research studies on human errors in RS image analysis is rather unexpected given that psychological research concerning image interpretation has existed since World War II. During the war, operators had to intently observe radar screens for several hours in order to detect whether an enemy was approaching. Despite the vital importance of their job, it was discovered that after a certain period of time, observers started missing signals (Parasuraman 1986). This drop in operator performance due to attention loss caused by performing the same monotonous tasks for long periods of time has been called *vigilance* and has been recognised and confirmed in a wide range of visual interpretation tasks. These include amongst others radiologists examining X-rays for traces of cancer (Laming and Warren

2000; Szalma 2006), luggage screening at airports (Bolfing *et al.* 2008) and industrial inspection tasks (Drury 1975).

These findings on sustained attention loss were the start of a fruitful line of research into operator performance that focused on three major aspects: (1) the way humans perceive things, (2) the individual (personal) differences that affect performance, and (3) the task-specific factors such as vigilance and training that have an impact on performance. Human perception is the organization, identification, and interpretation of sensory information in order to represent and understand the environment (Schacter *et al.* 2011). All perception involves signals in the nervous system, which in turn result from physical stimulation of the sense organs (Goldstein 2009). Although the senses were traditionally viewed as passive receptors, the study of optical illusions and ambiguous images has demonstrated that the brain's perceptual systems actively and pre-consciously attempt to make sense of their input (Gregory and Richard 1987). Optical illusions are thus a constant reminder of how human perception is a mental construct, influenced by context and prior knowledge, rather than a perfectly accurate camera-like registration of reality. For this reason insight in human perception is vital to understand the image interpretation process. Border delineation in RS image analysis is not something that one would immediately link with optical illusions. Nevertheless, Popple (2003) found that, when humans had to localise borders between two areas with different textures (horizontal and vertical), the border location was systematically biased towards the area where the texture was aligned with the border. In addition to such contextual effects, perception, and thus image interpretation, may also be influenced by inter-individual differences between operators. Bolfing *et al.* (2008) found several (visual) abilities that play an important role in the detection of dangerous objects in X-rayed luggage images at airports. Koller *et al.* (2009) on the other hand

found a negative correlation between age and both accuracy and reaction times in aircraft structural inspection. Szalma *et al.* (2006) confirmed that also attitudes (pessimistic or optimistic) may determine how an operator will respond to training for vigilance. Another individual factor potentially influencing image interpretation accuracy is search strategy. Every operator has his own strategy for screening an image varying from a random to a systematic approach. By monitoring the eye movements, operators can be trained to alter their strategy towards a more random or systematic search. Wang *et al*. (1997) observed that training in systematic inspection caused higher performance while training in random search had the opposite effect. Search time is normally followed by a decision moment where the operator has located the target, but needs to recognise, decide and react. Training can also have a major impact on this decision time, as a visual stimulus needs to be matched with representations in the visual memory (Koller *et al.* 2009). Also, people may differ significantly with respect to the information that they can retain in visual short term memory (Luck and Vogel 1997). If more information may be represented in visual memory, interpretation tasks become easier. Task-specific training can also be important to create more agreement on the definition of the studied object. Cooper *et al.* (2007) pointed out that in the delineation of tumors different experts expressed different ideas of what precisely represented a tumor.

Based on the long history of cognitive psychological research we assume that the overly confident belief in human judgment and interpretation of RS materials may not be justified. Given the high similarity between the above situations and the tedious, sometimes long-lasting routine tasks that are required from operators involved in RS image analysis, we propose that the insights from signal detection theory have been overlooked for too long in RS research and could be used to examine, understand, and

improve human performance in various image analysis procedures. To the best of our knowledge, no objective tool to evaluate the effect of operator performance on RS image analysis exists so far (Gardin *et al.* 2011). Therefore, we took a closer look at insights from basic psychological research on signal detection and employed these perceptivities to specific RS screening and interpretation tasks. Hence, our objectives were (1) to examine to what extent human performance in RS image analysis was liable to error; and (2) to assess which determinants were appropriate to explain inter-individual differences in performance. To this end, a number of experiments were run in which operator performance was examined as a function of time.

2. Method

2.1. Overview

As human image interpretation is highly variable among different operators, we aim to quantify operator variability and identify the human/external factors that potentially influence this variability. We collected data about human image interpretation variability by subjecting operators to a series of image digitization tasks. In order to address a large number of operators, data collection was carried out over the World Wide Web via a web application particularly designed for this purpose. Next, we collected data on human and external factors by inquiring operators about their personality through questionnaires and a visual memory test. Then, correlations were examined between digitization results and questionnaire results; and finally, a linear regression was performed to determine the influence of human and external factors on digitization performance. An overview of the method is shown in figure 1.

[PLEASE INSERT FIGURE 1 HERE]

The web application was developed in the C# language in an ASP.NET environment running on a Windows Server 2008. The interactive tests were developed using JavaScript and the maps were displayed with the open source JavaScript package Openlayers. All collected data was saved and processed in a PostgreSQL database with the PostGIS extension for geographical data. In figure 2 the site flow is schematically depicted.

[PLEASE INSERT FIGURE 2 HERE]

2.2. Selection of subjects

Operators were chosen in two different ways: (1) in view of fine-tuning and calibration, participants performed the interactive tests in a controlled environment (try-out); and (2) in order to ensure sufficiently large datasets, data collection was thence conducted via the World Wide Web (uncontrolled environment). In total 300 operators participated in the online experiment. Half of this group executed the test in a controlled environment (classroom setting). This group consisted of students and personnel of Ghent University (Belgium) who were either asked to complete the test as a practical exercise in a RS course (48) or who received a financial reward for their cooperation (95). The remaining part of the group consisted of people who voluntary finished the test online (157). In order to keep up the motivation to perform well, a prize was awarded to the best performing participant - the one who achieved the highest thematic and positional accuracies on the digitizing tasks. Nevertheless, there were also over 150 volunteers who started but did not complete the test.

2.3. Data collection

All participants were given a series of interpretation tasks and were also asked questions regarding their personality. Hence, for each participant, two types of information were collected: (1) performance on image interpretation tasks, and (2) a personal profile including human and external factors.

2.3.1. Image interpretation performance

For the image interpretation tasks we chose to assemble the experiment with a series of analogous digitizing tasks. This monotonous test was judged to be in line with the real-life working situation of a RS operator (e.g. localizing ground control points for image registration is mostly experienced as an unvarying and rather tedious task). Although more captivating to the operators, a higher degree of variability between the digitizing tasks would provide an unrealistic view on the time span over which they were able to keep their attention to the job. We examined operator performance (see Section 2.4.) as a function of time; hence a time limit was set to the interpretation of each image. This limit varied according to the specific digitizing task. We set a very comfortable time constraint, so there was no real time pressure to complete every single task. However, as in real working conditions, participants were obliged to stay focused and to keep up the pace of work.

For online digitization, the operators were offered two sets of aerial images, mainly selected in view of the availability of highly accurate reference data. The first image set was a series of aerial natural colour orthophotographs (Digital Mapping Camera (DMC), Ground Sampling Distance (GSD) =8cm) of the city of Ghent (Belgium) from which the Flemish Government inferred the GRB geographical identification database. GRB refers to Large-Scale Reference Database (in Dutch

Grootschalig Referentie Bestand) and consists of a digital land register containing highly detailed geographic information about specific ground objects (buildings, roads, etc.). The GRB is produced by a combination of photogrammetric interpretation of aerial very high resolution imagery and terrestrial techniques while following a strict process of accuracy assessment with checks on consistency, precision and completeness (AGIV, 2001). As the images provided in the web application offered much lower resolution, associated errors in the reference data were significantly smaller than the observed operator inaccuracies in the digitizing tasks of the online experiment. The specific assignments given to the participants were to digitize lamp posts (points), water bodies (polygons) and road networks (lines). The second set of images (Airborne Digital Scanner (ADS40), PAN GSD=5cm) featured the patchy olive vineyard landscape of Les Baux de Provence (France). The tasks here were to identify olive trees (points), to delineate olive parcels (polygons) and to digitize vine rows (lines). The reference data for these tasks were collected from different sources. Olive trees were georeferenced by locating every single tree with a GPS on the field; parcel boundaries were derived from the land register and the location of the vine rows were calculated based on the known distances of their specific planting pattern. Series of both image sets were alternately offered to the participants (table 1). All operators digitized at the same zoom-level: image zooming was not provided by the web application.

[PLEASE INSERT TABLE 1 HERE]

2.3.2. Personal profile

In order to identify a number of potential determinants for human performance in these tasks, we administered a comprehensive test battery of psychological variables, ranging

from personality and motivation to basic cognitive skills such as visual memory. The choice of these variables was inspired by the earlier psychological research as mentioned above (Bolfing *et al.* 2008; Koller *et al.* 2009; Szalma *et al.* 2006). More tests were not feasible as the participants were already subjected to a long effort. We also monitored external and technical factors (table 2) that could influence performance.

[PLEASE INSERT TABLE 2 HERE]

2.3.2.1. Human factors

The experiment started with a short list of questions related to *age*, *gender*, *digitizing experience* and *education*. Besides digitization experience, image interpretation skills were questioned as RS image interpretation can be a daunting task for people who lack experience. In order to obtain a preliminary idea of the participant's capability of interpreting RS imagery, a short assessment was conducted. The participants were presented with three images and four multiple choice questions about what they perceived. Their score was considered an indication for *interpretation experience*.

Next, the first real personality test was presented: a questionnaire measuring the five principal factors of human personality, also called the *Big Five* (McCrae and Costa 1987). The Big Five refers to the non-cognitive personality factors *agreeableness*, *conscientiousness*, *emotional stability*, *extraversion* and *openness*. Together, they comprise the most widely established and thoroughly validated theory of human personality structure (McCrae and Costa 1990). *Agreeableness* is the willingness to help other people, act in accordance with other people's interest and the degree to which an individual is co-operative, warm and agreeable. *Conscientiousness* is the preference for following rules and schedules, for keeping engagements and the attitude of being

hardworking, organized and dependable. *Emotional stability* encompasses dimensions such as being relaxed and independent. It addresses the degree to which the individual is calm, self-confident and self-restrained. *Extraversion* is the preference for human contacts, empathy, gregariousness, assertiveness and the wish to inspire people. Finally, *openness* measures the degree to which a person needs intellectual stimulation, change, and variety (Muller and Plug 2006; Borghans *et al.* 2008). These factors are reliable, stable, consistent across cultures, quite independent from intelligence, and they have been related to a wide range of human behaviour (Terracciano *et al.* 2005; McCrae and Terracciano 2005; McCrae *et al.* 2004).

Several studies have focused on the minimum length of a questionnaire to provide reliable results (Donnellan *et al.* 2006). We chose to present the participants with two questions presenting an overall view of each factor. As most explanatory power was expected from extraversion, emotional stability and conscientiousness, the list of questions was expanded with ten more questions on these factors. Participants rated the items on a five-point Likert scale, ranging from one (fully disagree) to five (fully agree). In order to assess the internal consistency of the questions, Cronbach's alpha (CA) was calculated for each factor.

For image interpretation tasks, the operators had to simultaneously process a large amount of visual information. Hence, the capacity of the visual memory plays an important role in the process. In addition to the Big Five, we included one cognitive variable: *short term visual working memory span*. Although everyday life is filled with a great deal of visual information, our short-term visual working memory can maintain representations of only three to four objects at a time (Luck and Vogel 1997; Xu and Chun 2006; Awh *et al.* 2007, Zhang and Luck 2008). For the present study, it is

important to note that visual working memory capacity is not a constant but instead varies considerably across individuals (Todd and Marois 2005; Vogel and Machizawa 2004; Vogel *et al.* 2005). To this end we added a widely accepted and reliable visual working memory span test to the online experiment. This test measures the number of objects that can be stored simultaneously in the visual working memory (Luck and Vogel 1997). Paired images with coloured blocks were presented to the participants. A first image with 4, 5, 6 or 7 blocks of different colours was displayed for 100 ms after which the image disappeared for 900 ms. Then the image was displayed again (1000 ms) with the same amount of blocks on the same locations, but in 50% of the cases, the colour of one of the blocks was changed (figure 3). The participant had to answer whether or not he/she saw the same image twice. As the images disappeared very fast, a couple of exercise images were presented for the participants to become acquainted with the concept. The real test subsequently consisted of 56 trials.

[PLEASE INSERT FIGURE 3 HERE]

The collected data consisted of the number of wrong and correct answers per size of array (i.e. number of blocks). Analogous to Alvarez and Cavanagh (2004), the visual working memory span of each participant was calculated as follows:

$$K = \frac{\sum_{S=S_1}^{S=S_2} S\,(H - F)}{(S_2 - S_1 + 1)} \qquad (6)$$

K measures the memory capacity, S is the size of array, H is the hit rate and F is the false alarm rate. S_1 and S_2 are respectively the minimum and maximum array size. The resulting value of K is compared to the sizes of the arrays. If an array size is smaller

than K or bigger than two times K, they are left out and a new K is calculated without these results. This iterative method is repeated until K remains constant.

Two other non-cognitive human factors that are known to strongly influence performance in these tasks are *motivation* and *comparative anxiety*. *Motivation* pertains to the participant's desire to obtain a good result. *Comparative anxiety* refers to the confidence that the participant has in his/her own abilities and performance and how much concern he/she puts in the performance of others. In this study, where the group of participants is very diverse and was motivated in different ways (payment, interest, willingness to support a scientific experiment or obligation), these two factors were expected to have a strong impact. Both human factors were surveyed through a questionnaire of ten questions that was presented after the image interpretation tasks. In this way, participants could make a statement about how they really felt during the online experiment instead of how they expected they would feel.

2.3.2.2. External factors

Finally, some questions were added to inquire about the circumstances in which the test was performed. For data acquisition in a controlled environment, this was not really an issue, but for the participants working over the internet, a large variability was expected in terms of *quality of computers/screens, amount of distraction, tiredness, time of day* and *amount of coffee* already consumed. Next to personal interpretation of screen quality by the participants, the characteristic *screen resolution* was also detected through a Java Script function and stored in the database.

2.4. Data analysis

Operator performance was quantified taking both thematic and positional accuracy into account. For thematic accuracy, the sensitivity index d-prime from signal detection theory (equation 1) was considered allowing for comparability with studies in other domains. Additionally, also measures commonly used in RS accuracy assessment were computed. The user's accuracy represented the correctness (equation 2), while producer's accuracy was rather a measure for completeness (equation 3) (Congalton and Green 2009, Matikainen *et al*. 2009). These two measures were also combined in one value, the $Mean\ Accuracy$ (equation 4).

$$d' = z\left(\frac{n_{R\,\&\,T}}{n_R}\right) - z\left(\frac{n_T - n_{R\,\&\,T}}{n_T}\right) \quad (1)$$

$$Correctness = \frac{n_{R\,\&\,T}}{n_T}100\% \quad (2)$$

$$Completeness = \frac{n_{R\,\&\,T}}{n_R}100\% \quad (3)$$

$$Mean\ Accuracy = \frac{2n_{R\,\&\,T}}{n_R + n_T}100\% \quad (4)$$

Where n_R is the number of elements that are present in the reference file, n_T is the number of elements digitized by the participant during the test, $n_{R\,\&\,T}$ is the number of hits (present in both the reference and the test image), and the function $z(x)$, $x \in [0,1]$, is the inverse of the cumulative Gaussian distribution.

Next to the issue of correct identification of the elements, also the precision of digitization was investigated. This positional accuracy was calculated as the mean distance between the test and the reference object (equation 5).

$$Mean\ Distance = \sum_{d=1}^{d=n} dn_d \quad (5)$$

d represents the radius of the buffer zone created around the reference units and n_d is the number of elements that were found within this buffer zone. Prior to analysis, performance outliers were filtered out by setting lower and upper thresholds of resp. $Q_1 - 1.5IQD$ and $Q_3 + 1.5IQD$ (with Q_1= lower quartile, Q_3= upper quartile and IQD = Inter Quartile Distance).

In order to gain clear insight in the group of participants and their working environment, we processed some basic statistics about their number, gender and age distribution, educational level, digitizing/interpretation experience, their personality, the working conditions, and the time they spent on the experiment. The latter allowed us to study vigilance, i.e. the drop of performance due to attention loss that potentially occurred after a longer period of digitization.

Finally, we determined the influence of human and external factors on digitization performance via correlation and regression analysis.

3. Results and discussion

3.1. Descriptive statistics of subjects

Figure 4 provides an overview of some basic statistics. As to gender, the total group consisted of 167 male and 133 female subjects. Figure 4(*a*) shows their level of education. Subjects who indicated high school or bachelor as their highest education level were mainly students who didn't complete their master studies yet. So, level of education should only be considered as the amount of education and not as an indication of IQ, although these two are obviously correlated (Lubinski 2004; Judge *et al.* 2010).

The high proportion of students is also reflected in the age distribution (figure 4(*b*)). The amount of time spent on the digitizing test is depicted in figure 4(*c*). As the time limit was not restrictive, participants were rather free in how fast they preferred to complete the test. This caused a large variability in the total time spent on the image interpretation tasks (20 to 150 minutes). However, the majority of the subjects worked in a time range from 40 to 80 minutes. In our research, speed is only considered as a factor that influences performance: subjects were simply and solely asked to perform as well as possible, and not to complete the test within a certain time span. In many working environments, speed is likewise considered a performance factor. The level of experience is shown in figure 4(*d*): 58 subjects indicated experience with digitizing tasks in the course of their professional career. There is a large group of subjects without any expertise, but most participants had at least basic digitization knowledge (169). Figures 3(*e*), (*f*) and (*g*) show the distribution of the personality factors (*e*); motivation and comparative anxiety (*f*) and visual working memory span (*g*). The personality factors, motivation and comparative anxiety all featured normal distributions and high internal consistency (CA$\in$[0.76, 0.91]); except for the factors agreeableness (CA=0.07) and openness (CA=0.49) where the lower consistency was expected given only two questions per factor. The normal distribution of visual working memory span indicated that the majority of the subjects were capable of simultaneously storing 2 to 5 objects in visual working memory, which is the typical range (Zhang and Luck 2008). Finally, figure 4(*h*) shows the scores for the factor *interpretation experience*: most subjects achieved a high score which was expected given the large majority of quite experienced subjects.

[PLEASE INSERT FIGURE 4 HERE]

Figure 5 shows factors of the test circumstances. Given the fact that the test was online, the large variability featured by the external factors was not at all surprising. Figure 5(*a*) demonstrates how the participants appreciated their working conditions. Even in the controlled environment, where test persons were subjected to the same circumstances, appreciation of the working situation strongly differed. We did not consider this as a flaw because a particular noisy working environment would be of influence or not, depending on how it was experienced by the participants: busy or not. One factor that might particularly cause interpretations flaws is screen quality: screen resolution is depicted in figure 5(*b*).

[PLEASE INSERT FIGURE 5 HERE]

After the completion of the experiments in the controlled environment we asked the participants to express their opinion about the test. Reactions varied strongly. Although the test was experienced as boring by most participants, many also indicated they did an effort to perform well at the beginning of the test, but motivation quickly diminished due to the amount of images they had to interpret. Another issue mentioned by many participants was how tiring the entire test was for the eyes (on average participants looked intently at the computer screen for 1h 20min 32sec to complete the entire test including the inquiry and digitizing tasks) and fingers (clicking on objects).

3.2. Human performance

The inter-individual differences in performance on the RS image interpretation tasks were quantified using the five accuracy measures: four quantifying thematic accuracy

(*Mean Accuracy* (%), *Completeness* (%), *Correctness* (%), and *d'* (-)) (table 3); and one measuring positional accuracy (*Mean Distance* (pixels)) (table 4).

[PLEASE INSERT TABLE 3 HERE]

[PLEASE INSERT TABLE 4 HERE]

As shown in table 3, performance was always far from 100%. Also, there was considerable variability across operators, in all six interpretation tasks. For instance, performance ranged from M = 11% to M = 100% for *Completeness* and from M = 35% to M = 100% for *Mean Accuracy* (with M indicating mean values). This illustrated that operators were not always the perfect interpreters that they were considered to be. Standard deviations were generally high indicating that performance values were spread out over a substantially large range. Also the magnitude of the errors made by the operators varied between digitizing tasks. For lamp posts for example, an average test person missed 20% of them whereas 16% of the objects they did detect weren't actually lamp posts. Furthermore, the lamp posts correctly identified were placed within an average distance of 5.43 pixels (i.e. 0.43 m) of the actual lamp post location (table 4). Trees on the other hand, seemed easier to detect correctly. The results proved conclusively that human operators were not infallible and that their performance both varied mutually and between the tasks they carried out. The latter might be related to the degree of complexity of the task at hand. Although the descriptive data were case-dependent and may not be extrapolated to other interpretation tasks, they clearly demonstrated inter-operator variability.

3.3. Group performance

With respect to professional background, the participants differed strongly. Both non-experienced operators and professional interpreters of aerial imagery participated in the test. In order to gain a clear insight into the impact of this subject variability, the performance of the different groups was compared. To also include the performance of the group of volunteers who dropped out before the end of the test, only the first half of the test was considered (figure 6). The performance distributions of volunteers, RS students and university personnel were very similar. The results of the students who joined the test because of a financial reward were considerably worse than of the other groups. This could have been caused by the fact that most of these participants didn't have any affinity with RS or GIS (students in psychology) and, contrary to the other groups, they had no real motivation to perform well (they were paid for just participating in the test, there was no particular incentive to perform well).

[PLEASE INSERT FIGURE 6 HERE]

3.4. Vigilance

In order to study vigilance (i.e. the drop of performance due to attention loss by doing the same monotonous tasks for a longer period of time), we focussed on the lamp post interpretation task. The offered image series within this task was the most extensive and thus facilitated exploration of the performance evolution over a longer time span. As not all participants were familiar with RS image analysis, and hence a learning factor could mask the tiring effect, the first series of DMC images were not considered in the analysis, but used to normalize the data per participant (normalization against mean

group performance). This was necessary to clarify the variability caused by human factors (as was established in previous sections).

Given the strong similarities with signal detection theory research, an apparent effect was expected for thematic accuracy, where, generally, performance remained consistent for a while and started decreasing rapidly after a certain breaking point. Figure 7 shows that performance first increased up to a certain maximum after which the expected drop occurred and performance steadily decreased. The initial learning effect was not surprising as participants were novices with respect to this specific task, and that some experience with the specific set-up and materials was necessary to achieve optimal performance. The subsequent drop in performance can be attributed to mental fatigue, confirming typical vigilance effects in earlier psychological vigilance research (Laming and Warren 2000; Parasuraman 1986; Szalma 2006).

[PLEASE INSERT FIGURE 7 HERE]

For experts as well as novices positional accuracy also showed a consistent decline over time (figure 8(a) and (b)). While the initial positional errors ($Mean\ Distance$) were about four to five pixels, at the end of the digitizing test these errors were in the magnitude of six to seven pixels. Concurrently variability increased considerably. An increase in positional error with half of the initial error size was remarkable for a short interpretation task which took most of the participants around one hour to complete. If a drop of performance could already be observed in an interpretation exercise of moderate duration, concerns could be raised about the working schedule of many operators where several hours without a break are no exception.

[PLEASE INSERT FIGURE 8 HERE]

3.5. Performance effect study

Based on tables 3 and 4, it was concluded that operator variability indeed was prominent. However, the question was raised whether and how much variability could be explained by human and external factors.

3.5.1. Correlation analysis

The effect of the human and external factors on operator performance was investigated using correlation analysis. The results for the lamp post digitizations are shown in tables 5 and 6). The correlations across all tasks are presented in table 7. Generally, the strongest correlations were found for the factors that were directly related to the test.

Participants who took more time (1) to localize the lamp posts generally performed better both thematically (C_{MA}=-0.27) and with respect to position (C_{PA}=-0.31) (table 5). Operators featuring a longer visual working memory span (2) also reached higher accuracy levels (C_{MA}=0.24; C_{PA}=0.18). Motivation (3) (C_{MA}=0.22; C_{PA}=0.21) and comparative anxiety (4) (C_{MA}=-0.19; C_{PA}=-0.24) displayed a strong respectively positive and negative correlation with performance. Additionally, the results suggested that men perform considerably better than women (5) (C_{MA}=-0.18; C_{PA}=-0.25). As women are known to be more patient in performing tedious and accurate jobs (Blatter *et al.* 2006; Feingold 1994), this might be an effect triggered by the test set-up, where men were more anxious to outperform their colleagues/classmates (Gherasim *et al.* 2013). Digitizing experience (6) (C_{MA}=0.18; C_{PA}=0.33) as well as interpretation experience (7) (C_{MA}=0.16; C_{PA}=0.26) contributed to

improved lamp post digitization results. As reflected by the high correlations with positional accuracy, experienced operators with fair RS image interpretation skills were particularly good at pinpointing the lamp posts. From the big five personality assessment, only extraversion (8) significantly influenced both thematic and positional accuracy. Generally, extraversion showed a negative impact on performance (C_{MA}=C_{PA}=-0.13). This negative effect could be explained by the difficulties people experience in keeping their attention to the task. Despite the fact that, based on psychological literature, we expected a strong positive link between conscientiousness (12) and accuracy (Shaffer and Postlethwaite 2013), conscientiousness also correlated negatively with performance (C_{MA}=-0.07; C_{PA}=-0.03). Emotional stability (11) on the other hand related positively to both thematic (C_{MA}=0.09) and positional (C_{PA}=0.15) performance. However, only the effect on positional accuracy was significant. Finally, seniors (14) were significantly better at localizing the lamp posts (C_{PA}=0.17) which was logical considering their level of experience.

[PLEASE INSERT TABLE 5 HERE]

The highly variable circumstances in which the image analysis was performed barely had an impact on operator performance (table 6). In essence, the subjective estimation of how busy the working environment was, showed significant correlations (C_{MA}=-0.16; C_{PA}=-0.23). Although the consumption of coffee did not affect thematic accuracy, it did influence the positional results (resp. C_{PA}=-0.13 and C_{PA}=0.13). Additionally, no performance differences were found between the groups who carried out the test in a controlled environment and volunteers working over the internet.

[PLEASE INSERT TABLE 6 HERE]

Table 7 shows the correlation results for all six digitizing tasks completed by the participant during the online test. Although various effects were playing at the individual task levels, we noticed that the human and external factors mainly influencing lamp post accuracy also affected thematic and positional accuracy of the other tasks, indicating that these factors generalized well across the digitizing tasks. The overall negative relations between conscientiousness and accuracy were unexpected considering similar psychological research (Shaffer and Postlethwaite 2013). As opposed to lamp post digitization, for the other tasks, age affected performance negatively. This effect was equally unexpected as seniors were assumed to perform better. With respect to external factors, only the level of distraction was of considerable relevance.

[PLEASE INSERT TABLE 7 HERE]

3.5.2. Regression analysis

Focusing on the lamp post objects as exemplar task and $Mean\ Accuracy$ as performance measure, we combined all human factors in a stepwise linear regression (table 8) and found that no less than 26% (R^2=0.26) of the variability in operator performance was explained by the human factors involved. In the perspective of psychological research, this was an adequate result considering the limited number of behavioural variables (Hemphill 2003). This equals about the amount of variance in future job performance explained by structured job interviews (Lievens 2011). When external factors were added to the model, 30% of the performance variability was

covered. This means that, by answering a limited number of questions and a small interactive test, a strong indication could already be provided as to the suitability of a candidate for interpreting RS images. As in our study the list of human factors was not exhaustive, an even stronger effect might be expected if intelligence measures would be included (which required too extensive testing for the present set-up).

[PLEASE INSERT TABLE 8 HERE]

4. Conclusion

Starting from the well-acknowledged finding in classical psychological research that human visual interpretation tasks are subject to high variability across time and individuals, the main purpose of this study was to quantify the variability in operator performance within RS image analysis, and relating this variability to external and individual human factors. Based on six different image interpretation tasks (digitization of lamp posts, water bodies, road networks, olive trees, olive parcels and vine rows), we found that operator performance was far from perfect, typically reached 80% and varied considerably across operators (with accuracy levels ranging from 11% up to 100%). These numbers illustrated the fact that operators are seldom the perfect interpreters they are supposed to be. We therefore encourage the RS community to seriously consider this issue as human interpretation results are still frequently used to benchmark automated mapping algorithms. Additionally, the amount of variability was dependent on the type of task presented to the operator: digitizing trees seemed to be less complex than localizing lamp posts. Given this high variability in performance, questions can be raised about the common practice of having RS image analysis carried out by only one operator..

Across the interpretation tasks, we found that operator performance was mainly determined by demographic, non-cognitive and cognitive personality factors, whereas external and technical factors influenced operator performance to a lesser extent. Performance was affected by the operator's desire to obtain good results, by the confidence the operator had in his/her own abilities and how much concern he/she put in the performance of others. Also, the operator's experience (both with digitizing tasks and RS image interpretation), how much time he/she was willing to spend, and how many representations he/she was able to retain simultaneously in short-term visual working memory played in important role. Ultimately, men performed better than women at most interpretation tasks. With respect to external factors, more noisy and busy working environments negatively influenced operator performance.

We demonstrated that the above mentioned human factors (together with the remaining human variables featuring minor influence) were responsible for no less than 26% of the inter-individual differences in operator performance. When the external factors, working environment, tiredness and screen quality were taken into account, this increased to 30%. These results highlight the importance of individual human characteristics in operator functioning and demonstrate their impact on operator performance. Moreover, the individual differences that are predictive of high performance on RS image analysis tasks were indicated. This could lead to the development of an assessment instrument that is able to identify and select individuals that dispose of the appropriate knowledge, skills, and abilities to perform image analysis tasks with a high level of accuracy for longer periods of time. On the basis of such an assessment instrument, research institutions, private and governmental organizations would be able to screen and identify individuals for remote sensing tasks, leading to more accurate outcomes and thus, lower costs.

Finally, a marked performance drop over time was assessed, both for experts and novices, suggesting that in operational conditions long-lasting image interpretation jobs without regular breaks should be avoided.

5. Further research

Next to human and external factors (constituting someone's personal profile), the impact of image-related factors on the performance of human operators in RS image analysis should be considered. Generally, interpreters are viewing images featuring very different characteristics like spectral and spatial resolution or tone and texture. Implementing all these factors in the same test would have assembled too many sources of variability, so it was elected to make this the scope of further research.

Acknowledgements

The Belgian Federal Science Policy Office (BELSPO) provided funding through the Research Programme for Earth Observation Stereo II (Contract Nr SR/02/121). AGIV is gratefully acknowledged for providing the images and reference data. The members of the WAVARS project steering committee provided useful comments. We also thank the anonymous reviewers for their valuable suggestions and the time they have invested into their review.

References

ALBRECHT, F., LANG, S. and HÖLBLING, D., 2010, Spatial accuracy assessment of object boundaries for object-based image analysis. *Third international conference on all aspects of Geographic Object-Based Image Analysis*, 29 June - 2 July, Ghent, Belgium.

AWH, E., BARTON, B. and VOGEL, E.K., 2007, Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science,* **18**, pp. 622–628.

BLATTER, K., GRAW, P., MUNCH, M., *et al*., 2006, Gender and age differences in psychomotor vigilance performance under differential sleep pressure conditions. *Behavioural Brain Research*, **168** , pp. 312-317.

BOLFING, A., HALBHERR, T. and SCHWANINGER, A., 2008, How image based factors and human factors contribute to threat detection performance in X-ray aviation security screening. In: *HCI and Usability for Education and Work*, Holzinger A. (Ed.), pp. 419–438 (Berlin: Springer).

BORGHANS, L., DUCKWORTH, A., HECKMAN, J. and TER WEEL, B., 2008, The Economics and Psychology of Personality Traits. *Journal of Human Resources*, **43**, pp. 972–1059.

COLWELL, R.N., 1997, *Manual of Photographic Interpretation* (American Society for Photogrammetry & Remote Sensing).

CONGALTON, R.G. and GREEN, K., 2009, *Assessing the Accuracy of Remotely Sensed Data* (Boca Raton: CRC Press).

COOPER, J. S., MUKHERJI, S.K., TOLEDANO, A.Y., BELDON, C., SCHMALFUSS, I.M., AMDUR, R., SAILER, S., LOEVNER, L.A., KOUSOUBORIS, P., KIAN ANG, K., CORMACK, J. and SICKS, J., 2007, An Evaluation of the Variability of Tumor-Shape Definition Derived by Experienced Observers from CT-Images of Supraglottic Carcinomas (ACRIN protocol 6658). *International Journal of Radiation Oncology*Biology*Physics*, **67**, pp. 972–975.

DONNELLAN, M.B., OSWALD, F.L., BAIRD, B.M. and LUCAS, R.E., 2006, The mini-IPIP scales: tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment*, **18**, pp. 192–203.

DRURY, C., 1975, The inspection of sheet materials: Models and data. *Human Factors*, **17**, pp. 257-265.

FEINGOLD, A., 1994, Gender Differences in Personality - A Meta-analysis, *Psychological Bulletin*, **116**, pp. 429-456.

FOODY, G.M., 2002, Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, **80**, pp. 185–201.

GARDIN, S., VAN LAERE, S.M.J., VAN COILLIE, F.M.B., DUYCK, W., ANSEEL, F., DE WULF, R.R. and VERBEKE, L.P.C., 2011, Remote sensing meets psychology: a concept for operator performance assessment. *Remote Sensing Letters*, **2**, pp. 251–257.

GHERASIM, L.R., BUTNARU, S. and MAIREAN, C., 2013, Classroom environment, achievement goals and maths performance: gender differences, *Educational Studies*, **39**, pp.1-12.

GOLDSTEIN, E.B., 2009, *Sensation and Perception* (Wadsworth: Cengage Learning).

GREGORY, R.L. and RICHARD, L., 1987, Perception. In: *Oxford Companion to the Mind*, Gregory, R.L. and Zangwill, O.L. (Eds.), pp. 598–601 (Oxford: OUP).

HEMPHILL, J.F., 2003, Interpreting the magnitudes of correlation coefficients, *American Psychologist*, **58**, pp. 78-79.

HOFFMAN, R.R., MARKMAN, A. and CARNAHAN W.H., 2001, Angles of regard: psychology meets technology in the perception and interpretation of nonliteral imagery. In: *Interpreting Remote Sensing Imagery: Human Factors*, Hoffman R.R. and Markman, A. (Eds.), Chapter 2, pp. 11–55 (New York: Lewis Publishers).

JUDGE, T. A., ILIES, R. and DIMOTAKIS, N., 2010, Are Health and Happiness the Product of Wisdom? The Relationship of General Mental Ability to Educational and Occupational Attainment, Health, and Well-Being. *Journal of Applied Psychology,* **95**, PP. 454-468.

KOLLER, S.M., DRURY, C.G. and SCHWANINGER, A., 2009, Change of search time and non-search time in X-ray baggage screening due to training, *Ergonomics*, **52**, pp. 644–656.

LAMING, D. and WARREN, R., 2000, Improving the detection of cancer in the screening of mammograms. *Journal of Medical Screening*, **7**, pp. 24–30.

LECKIE, D.G., GOUGEON, F.A., WALSWORTH, N. and PARADINE, D., 2003, Stand delineation and composition estimation using semi-automated individual tree crown analysis. *Remote Sensing of Environment*, **85**, pp. 355–369.

LIEVENS, F., 2011, *Human Resource Management* (Leuven: Lannoo Campus)

LILLESAND, T., KIEFER, R.W. and CHIPMAN, J.W., 2008, *Remote Sensing and Image Interpretation* (New York: Wiley).

LUBINSKI, D., 2004. Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "General intelligence, objectively determined and measured". *Journal of Personality and Social Psychology*, **86**, pp. 96–111.

LUCK, S. J. and VOGEL, E. K., 1997, The capacity of visual working memory for features and conjunctions. *Nature*, **390**, pp. 279–281.

MADDEN, M., JORDAN, T., MASOUR, J. and WANG, J., 2010, GEOBIA-based spatio-temporal changes and geovisualization within wildland urban interfaces. In *Third international conference on all aspects of Geographic Object-Based Image Analysis*, 29 June - 2 July, Ghent, Belgium.

MCCRAE, R. R. and COSTA, P. T., 1990, *Personality in adulthood* (New York: The Guildford Press).

MCCRAE, R.R. and COSTA, P.T., 1987, Validation of the 5-factor Model of Personality across Instruments and Observers. *Journal of Personality and Social Psychology*, **52**, pp. 81-90.

MCCRAE, R.R. and TERRACCIANO, A., 2005, Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology*, **88**, pp. 547-561

MCCRAE, R.R., COSTA, P.T., MARTIN, T.A., ORYOL, V.E., RUKAVISHNIKOV, A.A., SENIN, I.G. HREBICKOVA, M. and URBANEK, T., 2004, Consensual validation of personality traits across cultures. *Journal of Research in Personality*, **38**, pp. 179-201.

MULLER, G. and PLUG, E., 2006, Estimating the effect of personality on male and female earnings, *Industrial and Labor Relations Review,* **60**, pp. 3–22.

PARASURAMAN, R., 1986, Vigilance, monitoring and search. In: *Handbook of Human Perception and Performance*, *Vol.2, Cognitive Processes and Performance*,

Boff, J.R., Kaufmann, L. and Thomas, J.P. (Eds.), pp 41.1–41.49 (New York, Wiley).

POPPLE, A.V., 2003, Context effects on texture border localization bias. *Vision Research*, **43**, pp. 739–743.

SARMENTO, P., CARRÃO, H., CAETANO, M. and STEHMAN S.V., 2009, Incorporating reference classification uncertainty into the analysis of land cover accuracy. *International Journal of Remote Sensing*, **30**, pp. 5309–5321.

SCEPAN, J., 1999, Thematic validation of high-resolution global land-cover data sets. *Photogrammetric engineering and remote sensing*, **65**, pp. 1051–1060.

SCEPAN, J., MENZ, G. and HANSEN, M.C., 1999, The DISCover validation image interpretation process. *Photogrammetric engineering and remote sensing*, **65**, pp. 1075–1081.

SCHACTER, D.L., GILBERT, D.T. and WEGNER, D.M. (2011), *Psychology* (US: Worth Publishers)

SHAFFER, J.A. and POSTLETHWAITE, B.E., 2013, The Validity of Conscientiousness for Predicting Job Performance: A meta-analytic test of two hypotheses, *International Journal of Selection and Assessment*, **21**, pp. 183-199.

SZALMA, J.L., HANCOCK, P.A., DEMBER, W.N. and WARM, J.S., 2006, Training for vigilance: The effect of knowledge of results format and dispositional optimism and pessimism on performance and stress. *British Journal of Psychology*, **97**, pp. 115–135.

TERRACCIANO, A., ABDEL-KHALEK, A.M., ADAM, N., *et al.,* 2005, National character does not reflect mean personality trait levels in 49 cultures. *Science,* **310**, pp. 96-100.

TODD, J.J. and MAROIS, R., 2005, Posterior parietal cortex activity predicts individuals differences in visual short-term memory capacity. *Cognitive, Affective, & Behavioral Neuroscience,* **5**, pp. 144–155.

VOGEL, E.K. and MACHIZAWA, M.G., 2004, Neural activity predicts individual differences in visual working memory capacity. *Nature*, **428**, pp. 748–751.

VOGEL, E.K., MCCOLLOUGH, A.W. and MACHIZAWA, M.G., 2005, Neural measures reveal individual differences in controlling access to working memory. *Nature*, **438**, pp. 500–503.

WANG, M.J.J., LIN, S.C. and DRURY, C.G., 1997, Training for strategy in visual search, *International Journal of Industrial Ergonomics*, **20**, pp. 101–108.

XU, Y. and CHUN, M.M., 2006, Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, **440**, pp. 91–95.

ZHANG, W. and LUCK, S.J., 2008, Discrete fixed-resolution representations in visual working memory. *Nature,* **453**, pp. 233–235.

ZHU, Z., YANG, L., STEHMAN, S.V. and CZAPLEWSKI, R.L., 2000, Accuracy assessment for the U.S. Geological Survey Regional Land-Cover Mapping Program: New York and New Jersey Region, *Photogrammetric Engineering and Remote Sensing*, **66**, pp. 1425–1438.

List of figures

List of tables

Table 1. Object to digitize with corresponding image type, number of images and total number of objects, chronologically ordered as offered during the online experiment.

Object	Image type	Number of images	Total number of objects
Lamp post	DMC	35	75
Olive parcel	ADS40	3	44
Water	DMC	3	7
Road network	DMC	5	5
Vine rows	ADS40	1	35
Olive trees	ADS40	2	446
Lamp post	DMC	50	112
Olive parcel	ADS40	3	40
Road network	DMC	2	2
Vine rows	ADS40	1	35
Olive trees	ADS40	2	470

Table 2. Human and external factors investigated in the experiment.

Factor	Source
Human Factors	
Age	Form
Sex	Form
Digitizing experience	Form
Level of education	Form
Visual acuity	Form
Color blindness	Form
Speed	Calculated
Extraversion	Test
Emotional stability	Test
Conscientiousness	Test
Agreeableness	Test
Openness	Test
Motivation	Test
Comparative anxiety	Test
Visual working memory span	Test
Interpretation experience	Test
External Factors	
Level of distraction	Form
Tiredness	Form
Screen quality	Form
Coffee consumption	Form
Time of day	Form
Resolution monitor	Calculated

Table 3. Mean (M), Standard Deviation (SD), Minimum (Min), Median (Med), Maximum (Max) and Standard Error (SE) of the four thematic accuracy performance measures (N=300).

	Factor	M	SD	Min	Med	Max	SE
	***Completeness* (%)**						
1	Lamp Posts	80.09	7.19	61.22	81.12	93.88	0.42
2	Parcels	84.42	7.95	60.04	85.26	100	0.49
3	Water	63.29	10.11	36.51	65.70	91.65	0.63
4	Trees	90.41	9.43	53.14	94.04	98.98	0.59
5	Vine rows	71.82	20.13	11.00	77.41	100	1.21
6	Roads	66.50	12.37	32.27	67.92	90.17	0.73
	***Correctness* (%)**						
1	Lamp Posts	84.04	6.29	66.48	84.66	96.51	0.37
2	Parcels	90.33	6.94	71.80	92.98	100	0.42
3	Water	91.69	3.99	79.03	92.38	100	0.25
4	Trees	94.73	2.59	87.01	95.57	100	0.16
5	Vine rows	83.40	8.49	62.34	82.79	100	0.50
6	Roads	79.61	9.01	53.97	80.90	97.71	0.53
	***Mean Accuracy* (%)**						
1	Lamp Posts	81.77	6.51	61.97	82.83	94.30	0.38
2	Parcels	86.35	5.34	69.87	87.29	99.68	0.33
3	Water	74.02	7.55	53.54	76.48	85.41	0.47
4	Trees	92.26	6.18	68.83	94.79	97.56	0.39
5	Vine rows	76.52	13.94	35.04	78.46	100	0.85
6	Roads	72.02	10.45	41.37	73.07	90.74	0.62
	d'* (-)						
1	Lamp Posts	1.92	0.51	0.45	1.92	3.23	0.03
2	Parcels	2.42	0.48	1.01	2.51	3.26	0.03
3	Water	1.87	0.43	0.11	1.97	2.76	0.03
4	Trees	2.88	0.82	0.73	3.10	3.94	0.05
5	Vine rows	1.60	0.64	0.07	1.56	3.67	0.04
6	Roads	1.42	0.59	-0.05	1.41	2.77	0.03

* Higher *d'* values indicate that the signal can be more readily detected.

Table 4. Mean (M), Standard Deviation (SD), Minimum (Min), Median (Med), Maximum (Max) and Standard Error (SE) of the positional accuracy performance measure (N=300).

	Factor	M	SD	Min	Med	Max	SE
	Positional Accuracy **(pixels)**						
1	Lamp Posts	5.43	0.77	3.26	5.38	7.72	0.04
2	Parcels	3.21	0.35	2.39	3.19	4.13	0.02
3	Water	2.66	0.35	1.90	2.61	3.70	0.02
4	Trees	4.07	0.47	2.80	4.04	5.36	0.03
5	Vine rows	1.52	0.27	0.86	1.53	2.18	0.02
6	Roads	3.79	0.35	2.91	3.78	4.65	0.02

Table 5. Mean (M), Standard Deviation (SD) and Correlation of human factors with thematic performance (Mean Accuracy, C_{MA}) and positional performance (Positional Accuracy, C_{PA}) (N=300) (Lamps Posts only).

	Factor	M	SD	C_{MA}	C_{PA}
				Mean Acc.	*Pos. Acc.*
(1)	Speed	2.04	0.89	-0.27*	-0.31*
(2)	Visual working memory span (N=235)	2.10	1.23	0.24*	0.18*
(3)	Motivation (CA[§]: 0.91)	3.38	0.77	0.22*	0.21*
(4)	Comparative anxiety (CA: 0.76)	2.56	0.58	-0.19*	-0.24*
(5)	Sex (male=0; female=1)	0.44	0.50	-0.18*	-0.25*
(6)	Digitizing experience	1.55	1.57	0.18*	0.33*
(7)	Interpretation experience (N=162)	2.88	1.04	0.16*	0.26*
(8)	Extraversion (CA: 0.90)	3.17	0.69	-0.13*	-0.13*
(9)	Level of education	2.16	0.83	0.13*	0.19*
(10)	Color blindness	0.02	0.14	0.10	0.07
(11)	Emotional stability (CA: 0.91)	3.48	0.72	0.09	0.15*
(12)	Conscientiousness (CA: 0.87)	3.52	0.64	-0.07	-0.03
(13)	Visual acuity	0.51	0.50	-0.06	-0.01
(14)	Age	27.92	8.49	0.05	0.17*
(15)	Agreeableness (CA: 0.07)	3.28	0.66	-0.03	-0.05
(16)	Openness (CA: 0.49)	3.77	0.75	0.00	0.06

* Significant for p=0.05

[§] Cronbach's Alpha (CA)

Table 6. Mean (M), Standard Deviation (SD) and Correlation of circumstances with thematic performance (Mean Accuracy, C_{MA}) and positional performance (Positional Accuracy, C_{PA}) (N=300) (Lamps Posts only).

	Factor	M	SD	C_{MA}	C_{PA}
				Mean Acc.	*Pos. Acc.*
(1)	Level of distraction	2.29	1.03	-0.16*	-0.23*
(2)	Coffee consumption	1.73	0.91	0.04	0.13*
(3)	Time of day	3.17	1.31	-0.09	0.02
(4)	Resolution monitor (N=216)	-	-	0.06	-0.13
(5)	Tiredness	3.03	1.05	0.03	-0.01
(6)	Screen quality	4.15	0.85	-0.02	-0.04

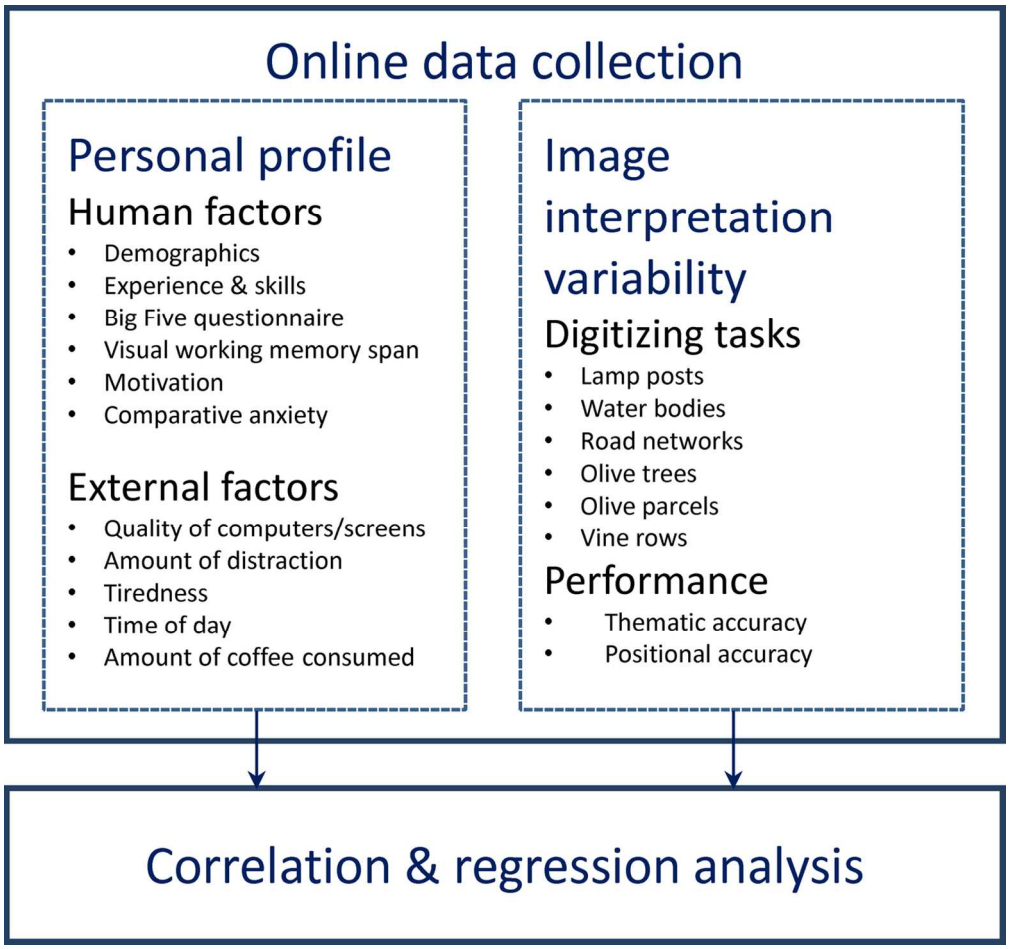
* Significant for p=0.05

Table 7. Correlation of human and external factors with thematic performance (Mean Accuracy, C_{MA}) and positional performance (Positional Accuracy, C_{PA}) (N=300) (all tasks).

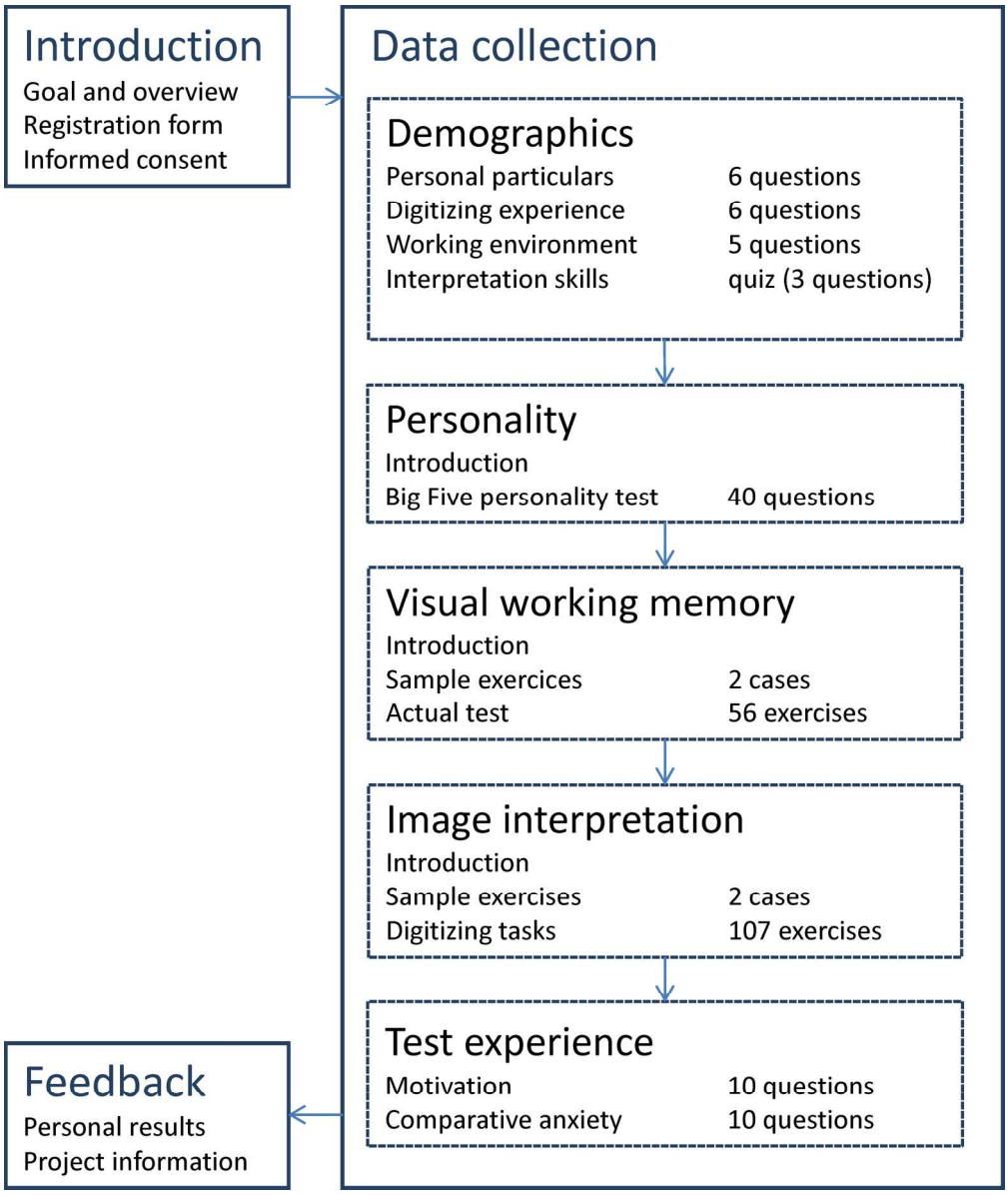
Factor	C_{MA}						C_{PA}					
Human	*Lamp Posts*	*Trees*	*Parcels*	*Water*	*Vine rows*	*Roads*	*Lamp Posts*	*Trees*	*Parcels*	*Water*	*Vine rows*	*Roads*
(1) Speed	-0.27*	-0.67*	-0.16*	-0.19*	-0.50*	-0.45*	-0.31*	-0.37*	-0.44*	-0.25*	-0.06	-0.39*
(2) Visual Working Memory Span (N=235)	0.24*	0.16*	0.15*	0.10	0.12	0.18*	0.18*	0.06	0.14*	0.02	0.02	-0.02
(3) Motivation (CA: 0.91)	0.22*	0.35*	0.06	0.20*	0.30*	0.27*	0.21*	0.24*	0.12*	0.10	0.05	0.19*
(4) Comparative Anxiety (CA: 0.76)	-0.19*	-0.09	-0.12*	-0.21*	-0.15*	-0.25*	-0.24*	-0.12*	-0.17*	-0.13*	0.01	-0.12*
(5) Sex (male=0; female=1)	-0.18*	0.12*	-0.09	-0.20*	-0.02	-0.23*	-0.25*	-0.03	-0.03	-0.14*	0.08	-0.15*
(6) Digitizing experience	0.18*	-0.10	0.10	0.21*	0.04	0.26*	0.33*	0.04	0.04	0.07	-0.14*	0.16*
(7) Interpretation experience (N=162)	0.16*	0.09	-0.08	0.19*	0.26*	0.25*	0.26*	0.07	0.10	0.20*	-0.08	0.15
(8) Extraversion (CA: 0.90)	-0.13*	0.02	-0.17*	-0.10	0.02	-0.04	-0.13*	-0.05	-0.02	0.05	0.01	-0.04
(9) Level of Education	0.13*	0.04	0.08	0.12*	0.03	0.21	0.19*	0.05	0.13*	0.07	-0.12*	0.10
(10) Color Blindness	0.10	-0.01	0.07	0.09	0.03	0.01	0.07	-0.17*	0.03	0.06	-0.06	0.06
(11) Emotional Stability (CA: 0.91)	0.09	0.11	0.02	0.07	0.03	0.16*	0.15*	0.08	0.02	0.04	-0.03	0.08
(12) Conscientiousness (CA: 0.87)	-0.07	-0.02	-0.12*	-0.12*	-0.07	-0.15*	-0.03	-0.01	-0.13*	-0.07	-0.01	-0.20*
(13) Visual Acuity	-0.06	0.06	0.04	0.08	-0.08	0.02	-0.01	0.17*	-0.01	0.07	-0.02	0.01
(14) Age	0.05	-0.12*	-0.13*	-0.01	-0.14*	-0.11	0.17*	0.04	-0.02	0.01	-0.17*	-0.13*
(15) Agreeableness (CA: 0.07)	-0.03	-0.02	-0.02	-0.03	0.00	-0.09	-0.05	-0.01	-0.09	-0.06	-0.06	-0.05
(16) Openness (CA: 0.49)	0.00	-0.07	-0.12*	0.03	0.04	0.00	0.06	0.02	0.04	0.04	-0.12*	-0.02
External												
(17) Level of distraction	-0.16*	-0.03	-0.03	-0.12*	-0.08	-0.19*	-0.23*	-0.03	-0.25*	-0.12*	0.09	-0.11
(18) Coffee consumption	0.04	-0.14*	0.06	0.12*	-0.06	0.00	0.13*	0.11	0.03	0.07	-0.09	-0.04
(19) Time of day	-0.09	-0.18*	-0.01	-0.08	-0.13*	-0.11	0.02	0.01	-0.08	-0.07	-0.03	-0.19*
(20) Resolution monitor (N=216)	0.06	-0.01	-0.11	-0.07	0.02	0.00	-0.13	0.11	-0.08	0.00	0.07	0.02
(21) Tiredness	0.03	0.15*	0.06	0.01	0.04	0.11	-0.01	-0.02	0.06	0.06	0.00	0.06
(22) Screen quality	-0.02	0.04	-0.09	-0.04	-0.03	-0.08	-0.04	0.00	-0.06	0.00	-0.01	-0.09

Table 8. Stepwise linear regression with Mean Accuracy as dependent variable (Step 1 to 5: human factors; Step 6: external factors).

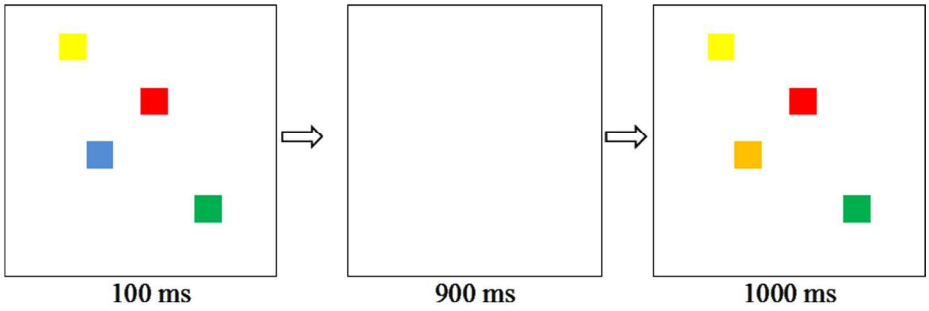
	Variable	b	SE(b)	t	p	R^2	ΔR^2
Step 1	Sex	-1.54	1.57	-0.98	0.33	0.06	0.06
	Age	-0.09	0.08	-1.05	0.30		
	Color blindness	-2.21	1.37	-1.61	0.11		
	Visual acuity	5.98	4.19	1.43	0.16		
	Level of education	0.37	0.97	0.38	0.71		
	Digitizing experience	0.84	0.69	1.22	0.22		
	Interpretation experience	0.67	0.74	0.91	0.36		
Step 2	Visual working memory span	1.40	0.57	2.46	0.02	0.12	0.06
Step 3	Agreeableness	0.58	1.03	0.56	0.58	0.14	0.02
	Openness	0.47	1.00	0.47	0.64		
	Conscientiousness	-1.89	1.12	-1.69	0.09		
	Emotional stability	0.45	1.07	0.42	0.67		
	Extraversion	-1.19	1.03	-1.16	0.25		
Step 4	Motivation	2.04	0.91	2.24	0.03	0.23	0.09
	Comparative anxiety	-0.38	1.26	-0.30	0.77		
Step 5	Speed	-1.67	0.72	-2.30	0.02	0.26	0.03
Step 6	Level of distraction	-0.01	0.71	-0.01	0.99	0.30	0.04
	Tiredness	1.27	0.77	1.65	0.10		
	Screen quality	-0.37	0.79	-0.46	0.64		
	Coffee consumption	0.23	0.76	0.31	0.76		
	Time of day	-0.87	0.55	-1.57	0.12		



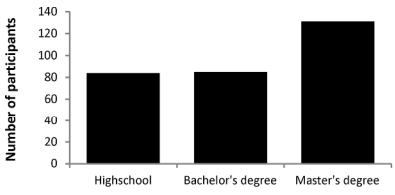
Method overview.
122x115mm (300 x 300 DPI)



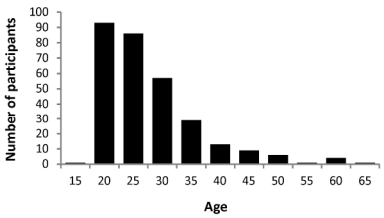
Site flow (http://wavars.ugent.be/wavarstest).
176x208mm (300 x 300 DPI)



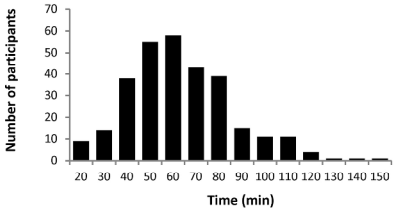
Visual working memory span test. Paired images with coloured blocks are presented for 100 ms to the participants. Then the image disappears for 900ms and reappears in another configuration: the number of blocks and their location remained the same, however, in 50% of the cases the colour of one of the block changed. Participants are queried whether or not they saw the same image twice.
154x59mm (300 x 300 DPI)



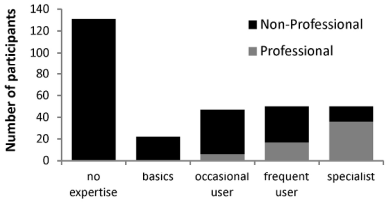
(*a*) Level of education



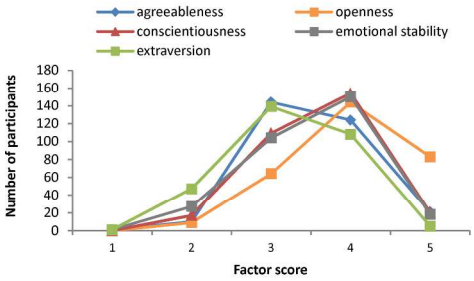
(*b*) Age distribution



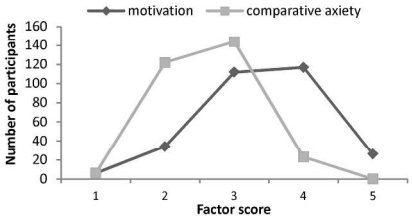
(*c*) Time to complete the digitizing tasks



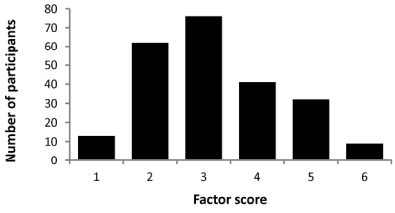
(*d*) Digitizing experience



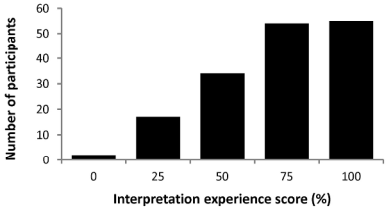
(*e*) Big five



(*f*) Motivation / Comparative anxiety

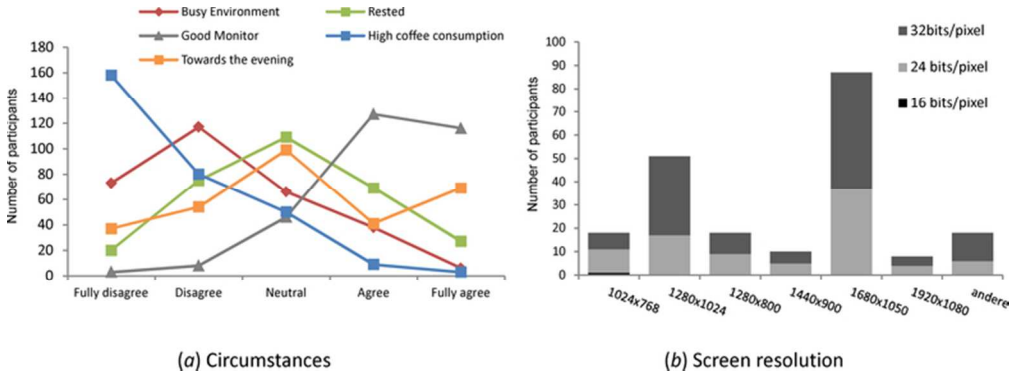


(*g*) Visual working memory span



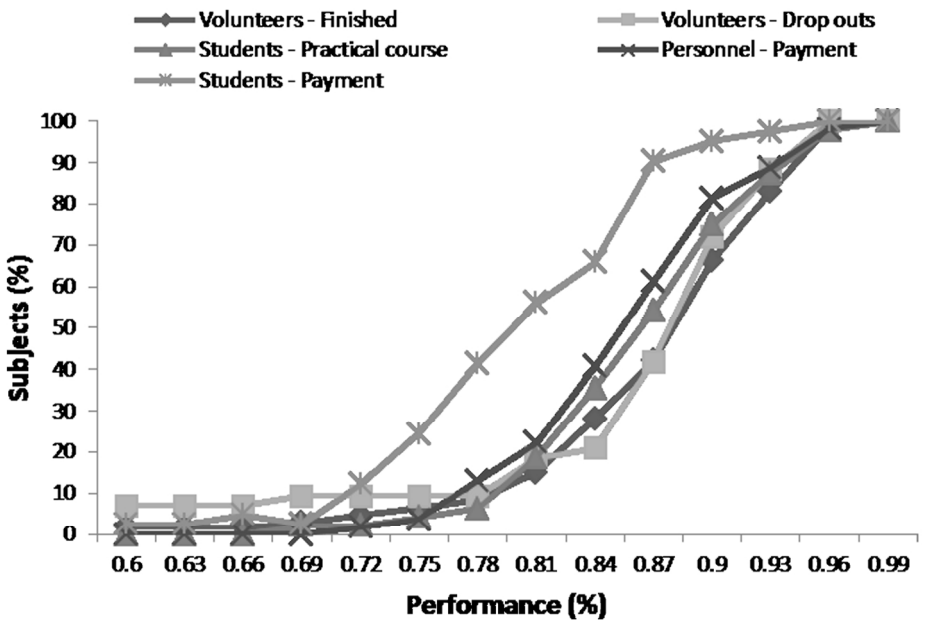
(*h*) Interpretation experience score

Descriptive statistics: human factors.
228x300mm (300 x 300 DPI)

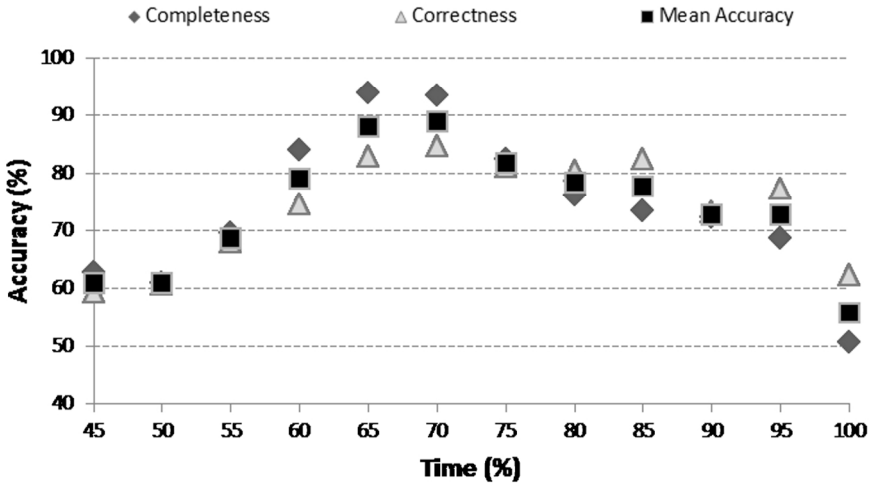


(*a*) Circumstances (*b*) Screen resolution

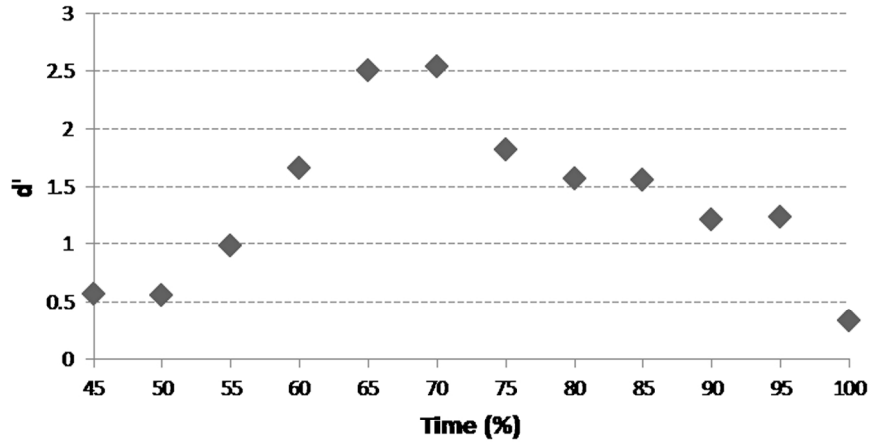
Descriptive statistics: external factors.
66x24mm (300 x 300 DPI)



Distribution of performance (%) per subject group for the first part of the test.
122x76mm (300 x 300 DPI)

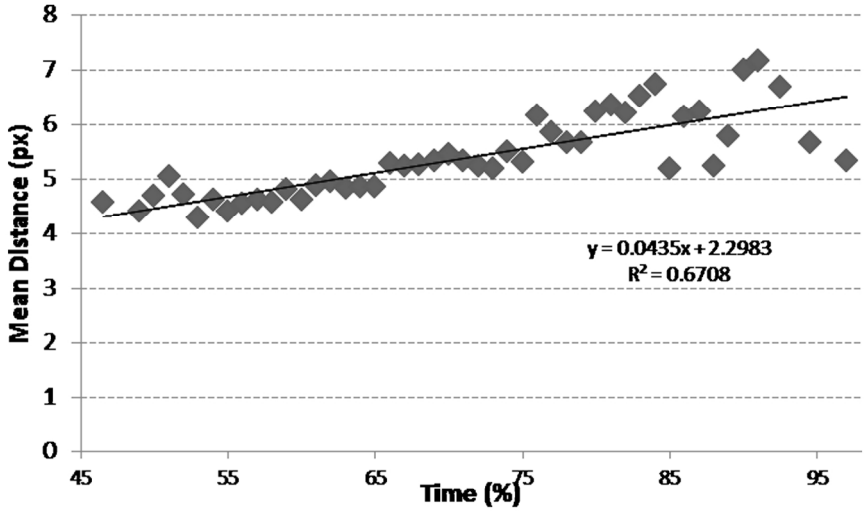


(a) Completeness, Correctness and Mean Accuracy

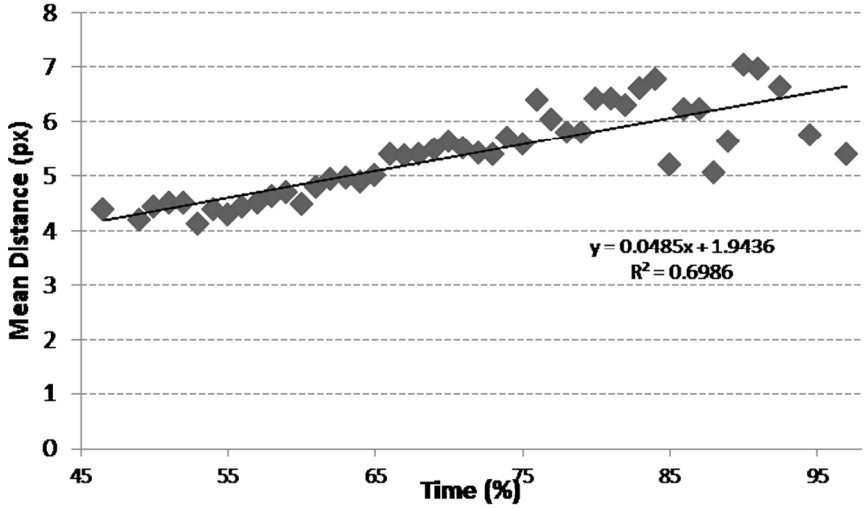


(b) d'

Thematic accuracy over time.
120x156mm (300 x 300 DPI)



(a) All participants



(b) Professionals only

Positional accuracy over time.
120x156mm (300 x 300 DPI)