From aircraft to e-government - using NASA-TLX to study the digital native's enrolment experience for a compulsory e-service

Chris Porter
Faculty of ICT, University of Malta
Malta

In recent years Malta launched a new e-service for students aged 16–18 who are applying for national exams. Adoption is compulsory and students also need to enrol for a national e-ID to gain access to the service. The e-service enrolment process is a pivotal part of the user experience, and without proper considerations it can become a major hurdle, stopping users from transacting online. This paper presents results from a two-stage study conducted with affected students to (1) measure and assess the impact of enrolment-specific design decisions on the students' lived experience (using NASA-TLX as a multi-dimensional and subjective workload assessment technique) and to (2) validate and critically assesses NASA-TLX's applicability and sensitivity in this context. This study gives particular attention to digital natives – people who have grown up with and are highly accustomed to digital technology (Prensky, 2001). This study shows that NASA-TLX is reasonably sensitive to changes in workload arising from various design-decisions within this context, however certain adoption caveats exist: (1) unsupervised NASA-TLX participants may provide significantly different results from supervised participants for most workload scales, (2) context-specific definitions and examples are necessary for most workload scales and (3) there are no major advantages arising from the adoption of a mean weighted workload (MWW) metric over raw TLX (RTLX).

Introduction and Aims

The enrolment process for any e-service can have a significant impact on the user's lived experience (Porter et al., 2012) and in turn on the success of the e-government service itself (Axelsson & Melin, 2012). In Western Europe the first generation of digital natives are starting to use e-government services. Most of these services require an online identity, which first-time e-government users have to create. The aims of this paper are to (1) develop an understanding on how enrolment-specific workload, as a multidimensional measure, impacts the digital native's experience with online services and (2) whether NASA-TLX is a suitable candidate to, in part, answer this question. Qualitative techniques are used to capture this citizen group's perceptions, expectations and reactions to identity-related tasks. This study also aims to determine whether NASA-TLX (1) is easy to understand and follow for younger (and untrained) participants, (2) whether it is applicable within this particular context (e-government) and (3) whether it is sensitive enough to detect changes in workload arising from different e-service enrolment process designs. Qualitative

In D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.) (2017). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

results will be treated as out of scope for this paper, and will not be presented nor discussed.

Background

Hart describes workload as 'the cost of accomplishing mission requirements for the human operator' (2006, par 1). The human costs in e-government include the citizen's inability to use an e-service - which could in turn result in sanctions, such as a fine for not paying a congestion charge on time (Inglesant & Sasse, 2007), or loss of opportunities, such as having to use otherwise productive time to visit a government department in person. However, the risk is also on the service owner: if the human cost for a service is such that e-services are not used, the government will also have to absorb costs for handling that particular transaction via traditional channels. This can also have political ramifications in that a negative experience will generally reflect negatively upon the government's image of efficiency and competence.

Cain argues that different workload measurement techniques actually assess different aspects of workload and this heterogeneity of focus stems from the 'lack of an accepted definition of workload' (2007, pg 7). According to the author different people have different perspectives on the meaning of workload, including (1) the task demands imposed on the user, (2) the effort the user needs to make to satisfy such demands and (3) the consequences of attempting to meet such demands. Sasse et al. (2014) adopted the GOMS-KLM approach (Goals, Operators, Methods and Selection rules – Keystroke-level Modelling) to assess the workload imposed by authentication events in terms of the time taken for a user to complete them. GOMS-KLM, introduced by Card et al. (1980), evaluates workload by deconstructing tasks into a set of basic actions, or steps, on which time measurements are taken. Although it is an important benchmarking technique to help practitioners determine the best and worst case scenarios in terms of user performance and effort for a given task, its simplicity might deter from its potential to provide measurable information on aspects such as frustration and self-confidence which, from a lived experience perspective, are also important considerations for the design of better security mechanisms. The author believes that the time taken to fill in a form does not necessarily imply a negative user experience, especially if the benefits obtained from using the e-service offset the cost associated with accessing it. For instance, a tax return e-service requiring users to authenticate by selecting a digital certificate, submitting a one-time password and filling in several other fields might still be worth the while for a professional who would otherwise need to regularly fill in and post paper-based forms on behalf of his clients. Workload can affect users in different ways, and for different reasons, and this impact may also vary across contexts of use.

For this reason, the author turned his attention to NASA-TLX – a multi-dimensional and subjective workload assessment technique. While developing NASA-TLX, Hart and Staveland (1988) examined ten workload-related factors, retrieved from sixteen experiments. Six of these factors were then proposed as a multi-dimensional rating scale combining magnitude and source information 'to derive a sensitive and reliable estimate of workload' (Hart and Staveland, 1988, pg 139). This was accomplished

after a series of statistical tasks, mainly to determine the sensitivity of each factor on workload. In NASA-TLX, physical and mental workload are also measured along with cognitive workload. This technique was originally developed for use in aviation flight-deck design; however, now it has been widely adopted for alternative uses and is also being used as a benchmark against which other workload measuring techniques are evaluated. Rubio et al. (2004) surveyed a number of studies which adopted subjective workload (cognitive) rating techniques. The authors ranked NASA-TLX at the forefront of sensitivity to experimental changes in workload conditions. This is also confirmed in Garteur's *Handbook of Mental Workload Measurement* (Garteur, 2003). Hill et al. (1992) also rated NASA-TLX as the most sensitive to workload changes, followed by MCH (Modified Cooper-Harper) and finally SWAT (Subjective Workload Assessment Technique).

NASA-TLX allows subjects to record data post-task, and thus certain physiological and time span-dependent effects may be in conflict to what is recalled by the subject. Techniques to counteract this issue include (1) screen-recording playback and (2) video-recording playback of the tasks performed. These techniques are designed to facilitate retrospective workload rating (Garteur, 2003). NASA-TLX uses six workload factors, or dimensions, and measures their relative contribution in influencing the user's perceived overall workload. Twenty years after presenting NASA-TLX, Hart (2006) reviewed the current state of the technique. It was found that most recent studies using this technique handled investigations on interface design and evaluation, with 31% focusing on visual and auditory displays and 11% on input devices. Seven percent of the studies were carried out with users of personal computers. Hart notes that NASA-TLX can be used in various situations, from aircraft certification to website design. This study proposes the use of NASA-TLX to measure enrolment-specific workload, primarily because of its multi-dimensional nature and overall performance sensitivity. Various other advantages of NASA-TLX include: ease of use; practicality of the method; reduction of between-rater variability (due to the adoption of weighted rankings) and the availability of clear instructions, supporting tools and case studies.

Study Context

The examinations department stipulated that students are to use a new e-service to register for their A-level examinations. Unless there were exceptional circumstances, students could not apply via the traditional method of visiting the examinations department in person. A 'Click Here to Apply' button was made available on a clean and easy to follow landing page at https://exams.gov.mt. Once clicked, students were asked to login using their e-ID credentials. No immediate information is given on how to obtain an e-ID. Instructions on how to enrol for an e-ID were provided in another e-government page, and at the time the process consisted of the following steps:

1. Visit the registration office in person (on average it takes 30 minutes each way by bus)
2. Go through a short enrolment process (on average it takes 5 minutes to complete and students need to present their national ID card and a valid

email address). Queues are possible since this is a central-government office
3. Receive a security PIN by post at the address given at enrolment
4. Activate the e-ID account using the PIN received by post and a password received at the email provided in step 2
5. Create a new password that adheres to a strict password policy

Once students are successfully enrolled on the National Identity Register, they are able to proceed to register and pay for their A-level examinations through the e-service website at https://exams.gov.mt/.

Method

Participants

Two sets of participants were involved in this study, one for each of the two phases discussed below, namely the (1) collection and analysis of users' experiences via an online questionnaire and the (2) follow-up workshops to verify and validate NASA-TLX ratings.

Process

The author's goal was to capture as much feedback as possible from the pool of students sitting for their exams. An online questionnaire was opted for since it would help (1) reach as many students as possible while (2) minimising disruptions to their studies. A number of interesting insights and recommendations emerged during this exercise. It was also felt that this study would benefit highly from a second intervention through which the initial results could be validated and substantiated. This was the motivation for the second part of the study which offered the opportunity to assess the applicability and understandability of NASA-TLX with digital natives and to investigate its sensitivity towards workload induced by enrolment-specific factors. Students who indicated that they would be willing to participate in follow-up meetings were contacted and a series of five workshops were scheduled. All ethical considerations recommended by the research ethics committees at the respective institutions were observed for both phases of the study.

Results

Three data sets were generated following this study: (1) qualitative results from the questionnaire outlining experiences for the various subgroups, (2) quantitative workload data obtained from the questionnaire's NASA-TLX assessment and (3) data from follow-up sessions which includes both qualitative and TLX related information. Thematic insights arising from the transcribed qualitative data will not be presented here.

Unsupervised NASA-TLX – online questionnaires

The questionnaire was sent to over 1000 students who were sitting for their A-Level examination sessions. A total of 134 valid responses were received (13% response

rate). Sixty-two percent of the participants were female, 21% male and 17% decided not to disclose their gender. Eighty-one percent of students declared that they fall within the 16–18 age-group, while 15% chose not to disclose their age. Four participants stated that they are aged 19-24, and one was over 25 years of age. Only those falling within the 16–18 age bracket were considered in the analysis stage. Furthermore, around 10% (13) of the respondents accepted the invitation to participate in one of a series of follow-up workshops held in the following months.

The second part of the questionnaire was an online version of the TLX workload assessment procedure. Initially students were asked to rate the six sub-scales (or workload dimensions) for the exam registration task (including e-ID enrolment if applicable), followed by the pairwise comparison to get a weighted overall workload measure (mean of weighted ratings). The six workload dimensions are Mental Demand (MD), Physical Demand (PD), Temporal Demand (TD), Own Performance (P), Effort (E) and Frustration (F). The overall task load index (MWW) was calculated for each participant, and averaged across the various student subgroups (see Figure 1). The cohort who used the e-service to complete the task, provided an overall mean weighted workload (MWW) of 42 (± 18.59) while those who registered for their exams in-person reported an overall MWW of 57 (± 14.11). These values, and particularly the variance in the online task's MWW, are not enough to draw any reliable conclusions on the users' experience. It would therefore be necessary to drill down into the various sources of workload while also analysing the process through which these values have been produced.

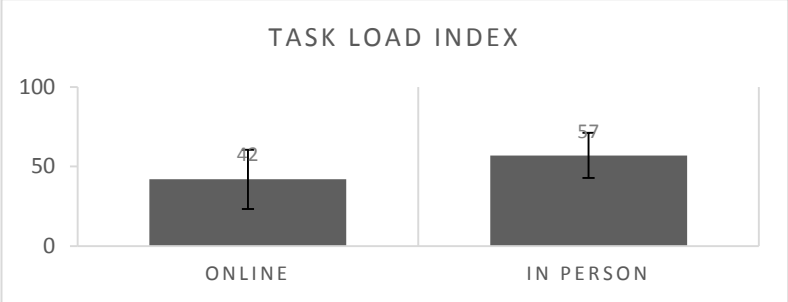

Figure 1. Mean weighted workload (MWW) for e-service users (online) and for those who adopted the offline exam registration process (at the exams registration department).

The average rating for the online method takes into consideration the ratings given by students who already had an e-ID and also by those who had to enrol for one. Table 1 shows how students who already owned an e-ID weighted the different workload dimensions.

Table 1. Workload dimension weighting by students who used the e-service and who already owned an e-ID

	MD[1]	*PD*[2]	*TD*[3]	*OP*[4]	*E*[5]	*F*[6]
Mean	3.7	0.4	1.9	2.6	2.2	4.3
Median	4	0	2	3	2	4
Std.Dev	1.2	0.9	1.1	1.4	0.9	0.7

Table 2. Workload dimension weighting by students who used the e-service but had to enroll for an e-ID

	MD[1]	*PD*[2]	*TD*[3]	*OP*[4]	*E*[5]	*F*[6]
Mean	2.8	1.4	3	2.1	2.2	3.7
Median	3	1	3	2	2	4
Std.Dev	1.4	1.5	1.4	1.4	1.1	1.3

Adjusted ratings are obtained by combining these weighted dimension values with raw ratings, as shown in Figure 2. In this case, Physical Demand is the lowest contributor to workload (adjusted rating = 2.5) however Frustration has an adjusted rating of 247, making it the highest contributor. Mental Demand follows Frustration, and thus these have a great influence on the average overall MWW. On the other hand, Table 2 shows how students who had to enrol for an e-ID weighted the different workload dimensions (out of 5). Figure 3 shows the respective adjusted ratings for this group. At a glance it is evident that this group of students had a different experience than the previous group and reported an increase in Physical and Temporal Demand. Frustration is still the highest contributor to workload, given an average weighting of 3.7, followed by Temporal Demand (3).

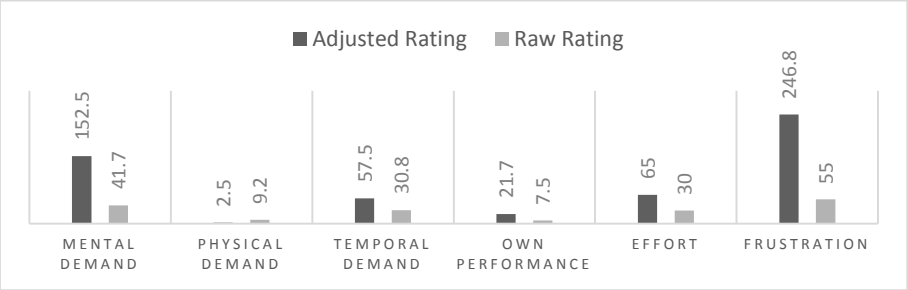

Figure 2. Adjusted rating for e-service users who already owned an e-ID (adjust rating = workload dimension weighting x raw rating)

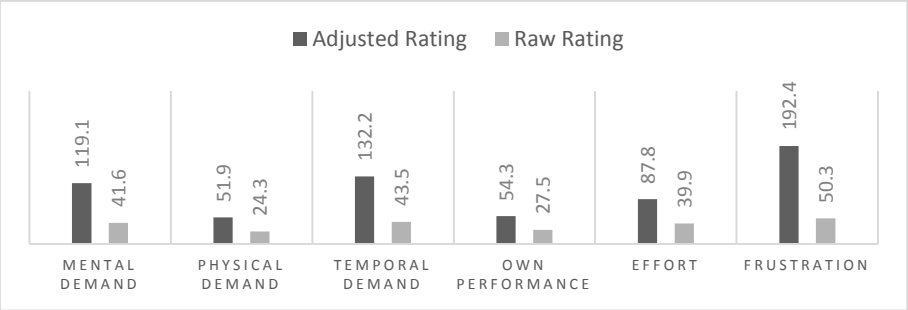

Figure 3. Adjusted rating for e-service users who had to enroll for an e-ID (adjust rating = workload dimension weight x raw rating)

Given this information, it can be seen that both groups of students (those who already had an e-ID and those who had to enrol for one) exhibited high levels of workload, albeit, for different reasons:

- *Those who had an e-ID:* Overall Task Load Index (TLX) was high mainly due to Frustration and Mental Demand. Causes for this outcome were various, including lack of process clarity, preference for traditional means, lack of trust in online systems and site performance.
- *Those who did not have an e-ID:* Overall TLX was high due to Frustration, Temporal and Mental Demand. Causes for this outcome were various, mostly due to the hassle involved to get an e-ID (e.g., waiting time at the e-ID enrolment office). Physical Demand was also significantly higher than that reported by the previous subgroup.

Supervised NASA-TLX – follow-up workshops

Students who agreed to participate in follow-up sessions were first asked to discuss their experience with the exam registration process and compulsory e-ID enrolment. Following this they were asked to compare and rate the perceived effort required to enrol for various online services including social networks, e-learning tools, payment gateways, email services and e-commerce sites. Each group had to reach a consensus for each rating decision and their interaction was observed. Following this, students were asked to go through a set of nine fictitious enrolment processes upon which workload measurements were taken. In all, 13 students agreed to participate in a series of follow-up sessions in small groups, eight of whom were female and five of whom were male. Their median age was 17 years old. All participants had just finished their A-level examinations

Perceptions on workload for popular online services

Before delving into the supervised NASA-TLX exercise, it was decided to conduct a series of semi-structured group-discussions without the use of rigid workload measurement techniques. This allowed for a consensus-driven thought process on the concept of workload as well as merits and de-merits of different enrolment processes adopted in popular online services. Each group of students (of 2 to 4 participants) was presented with a list of online services that they might have used at any point in time (e.g., Gmail, Facebook, Skype, PayPal and Hotmail amongst others). The most commonly used services for each group were then listed on a white board next to a rating scale indicating the level of perceived effort required to enrol for that specific service (i.e., easy, medium, difficult/annoying).

Figure 4. Participants had to agree on the level of perceived enrolment-specific workload (from personal experience) for several online services

First, students discussed the elements in enrolment they thought were responsible for workload from an individual perspective (this data will not be presented in this paper as it is deemed to be out of scope). Furthermore an agreement had to be reached on the relative level of perceived workload for each services' enrolment process in relation to other services' (as a group). Both mean and median values for the most commonly used services across all groups are presented in Figure 5. Feedback provided by different groups was normalised according to each group's rating patterns; that is, some groups always rated high, while others were more conservative. This made it possible to generate high-level, cross-group observations. Table 3 adds some context to these scores, providing annotations for the respective services' enrolment processes.

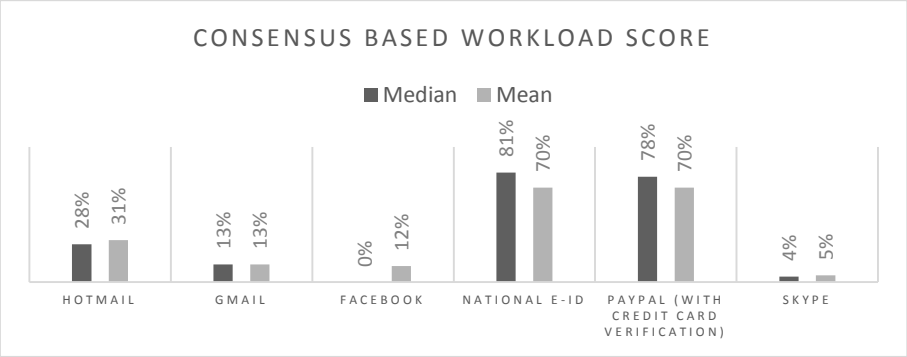

Figure 5. Perceived enrolment-specific workload for the most common online services

In a previous qualitative investigation (see Porter et al., 2013), it was established that; Items to Recall (ItR), Items to Generate (ItG), Interruptions to daily routines (I) and Delays (D) are central themes when it comes to sources of workload within enrolment processes. ItR represents the number of fields a user has to fill during the enrolment process. ItG represents a measure of the number of secrets the user has to

come up with (e.g., PIN, password). Any major interruption necessitating the user to go out of her way to complete the task is represented through I (e.g., visit an office to complete the process). Finally, D represents any form of interruption which introduces a delay in the process itself, but without disrupting the users' daily routines (e.g., a minor delay is introduced when an activation email is sent to the user, whereas a major delay is introduced when the service provider requires a day or two to conduct manual verification on submitted data).

Table 3. Various services' enrolment processes, their design factors and consensus based perceived workload

	ItR[1]	*ItG*[2]	*I*[3]	*D*[4]	*Perceived Workload (by consensus)*
Hotmail	10	2	No	No	28%
Gmail	8	2	No	No	13%
Facebook	6	1	No	No	0%
National e-ID	NA	3	Yes	Yes	81%
PayPal[5]	13	1	Yes[6]	Yes[7]	78%
Skype	11	2	No	No	5%

[1]Items to Recall [2]Items to Generate [3]Interruptions to daily routines [4]Delays [5]Including credit card verification [6]User needs to get hold of a bank statement [7]Can take several days until transaction is visible in a credit card/bank statement

Sensitivity of NASA-TLX

During the follow-up sessions, students were also individually asked to go through a number of fictitious enrolment tasks for fictitious e-services. These tasks are based on common enrolment process configurations generally used in e-government services. A number of e-services from around the world were surveyed and for each service's enrolment-process the researcher recorded its ItR, ItG, I and D values. This afforded the researcher the possibility to construct a set of fictitious tasks, based on real-world services with increasing levels of identity assurance requirements and workload (see Table 4). Table 5 shows how these nine tasks map onto real-world e-services.

Table 4. Set of nine enrolment tasks generalised from a survey of commonly found design configurations across various e-services (from low to high workload and assurance levels)

Design factors	**Fictitious enrolment tasks**								
	Low workload			*Medium Workload*			*High Workload*		
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
ItG	0	1	1	2	2	3	4	3	3
ItR	1	2	5	4	5	6	6	9	NA
D	No	No	Minor[2]	Major[3]	Major[4]	Minor[5]	No	Minor[6]	Major[8]
I	No	No	No[1]	No	Yes[4]	No	No	No	Yes[7]

[1]Credit card details are required [2]Wait a few minutes for activation email [3]Wait three days before account is activated [4]Visit closest outlet to confirm identity [5]Upload a recent photo [6]Call free-phone to activate account [7]Visit enrolment office during specific opening hours [6]Three day waiting period till an activation PIN is received by post

Table 5. Examples of real-world e-services adopting enrolment processes similar to the ones presented in Table 4 (as at 2013)

Task	*Based on…*
A	Directorate of labour (Iceland)
B	Estonian e-government portal (Estonia)
C	Birth certificates (Ontario)
D	Comune di Milano (Italy)
E	Student finance (England)
F	Study permits (Canada)
G	Inland revenue (Italy
H	Access key registration (Canada)
I	e-ID registration (Malta)

For each task, a NASA-TLX evaluation was carried out. It was decided to maintain the final NASA-TLX pairwise rating and thus generate a weighted workload rather than a raw TLX score (see Table 6 for resulting weighting values). It is evident from the weighting exercise that digital natives consider Frustration (F), Physical Demand (PD), Temporal Demand (TD) and Effort (E) as the major sources of workload (in this order). Frustration (F) was presented as a measure of irritation, stress and annoyance during the task while Effort (E) was explained to be the level of mental and physical work required to accomplish the task. This corroborates with the consensus based perceived workload levels shown in Table 3 whereby the highest workload scores were given to those enrolment processes that interrupted the primary task. In the National e-ID case students had to visit an office in Valletta, while in PayPal's case participants had to wait a couple of hours or days until a small PayPal transaction was processed and made visible on the credit/debit card statement. The transaction details on the statement contain an activation code which is required to complete the verification process (i.e., to confirm card ownership).

Table 6. Workload dimension weighting by students following the final pairwise comparison

	MD[1]	*PD*[2]	*TD*[3]	*OP*[4]	*E*[5]	*F*[6]
Mean	0.7	3	2.7	1.7	2.5	4.4
Median	1	3	3	1	3	5
Std.Dev	0.8	1.4	1.1	1.2	1	1

The participants' overall weighted workload values for each of the nine fictitious enrolment processes presented during this session are shown in Table 7.

Table 7. Median value for the mean weighted workload (MWW) score across all participants for the nine fictitious enrolment processes

Task	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
MWW	0%	0%	18%	11%	32%	14%	12%	21%	81%

A series of tests using the Related-Samples Wilcoxon Signed Rank non-parametric test for non-normally distributed data were carried out to determine whether there is a statistically significant difference between reported workload levels and corresponding tasks designed to be incrementally demanding. The null hypothesis set for these tests is that no statistically significant increase in perceived workload exists between tasks that are designed to be incrementally demanding. In some cases, although the task was intended to be less demanding than the subsequent one, it turned out that digital natives perceived it as more demanding; although a fairly low statistical significance is reported. For example, tasks C and D as well as tasks F and G whereby the null hypothesis was retained.

This can be explained by referring to the participants' supervised workload dimensions' weighting values (see Table 6) wherein Physical Demand (PD) and Temporal Demand (TD) (both given a weight of 3) are considered to be two major contributors to workload, as opposed to Mental Demand (MD) (weight of 1). Although tasks C and F are less demanding than their subsequent tasks (D and G respectively), with lower levels of mental demand (MD), they present users with more physical (PD) and temporal demands (TD) (i.e., travelling, looking up information and waiting for account activation).

Table 8. Tests to determine whether there is a statistically significant difference between reported workload levels for tasks designed to be incrementally demanding

Null Hypothesis[1]	*Significance (.05)*	*Decision*
PEW[*] for Task C over Task B	.001	Reject the NH
PEW for Task H over Task G	.033	Reject the NH
PEW for Task I over Task H	.003	Reject the NH
PEW for Task C over Task A	.001	Reject the NH
PEW for Task F over Task E	.039	Reject the NH
PEW for Task H over Task B	.001	Reject the NH
PEW for Task G over Task C	.039	Reject the NH

[1] **Null Hypothesis (NH):** The median of differences between each pair of data sets is equal to 0 (i.e., there is no statistically significant increase in perceived workload for subsequent incrementally demanding tasks).

[*] *PEW*: Perceived Enrolment Workload

Supervised vs unsupervised NASA-TLX

Consider Tables 1, 2 and 6. The weighting values for some of the workload dimensions provided via the online questionnaire (unsupervised) are considerably different from those provided for the same dimensions during the follow-up sessions (supervised) – see Table 9 for a synthesis of results. This presents the possibility that participants who had no immediate supervision, as opposed to the supervised group, may have interpreted the rating scales differently from the supervised group. If this is the case, the unmodified (original) NASA-TLX process would not be suitable in an unsupervised environment and with untrained participants. A set of tests are presented below to assess this hypothesis

Table 9. Workload dimension weighting (median) varied when students were supervised as opposed to unsupervised responses (i.e., no immediate help was available)

	MD[1]	*PD*[2]	*TD*[3]	*OP*[4]	*E*[5]	*F*[6]
Unsupervised (online)	4	0	2	3	2	4
Unsupervised (online without e-ID)	3	1	3	2	2	4
Supervised (follow-up sessions)	1	3	3	1	3	5

Given a non-normal distribution for the workload dimensions' weighting, a set of non-parametric tests were conducted using the Related Samples Wilcoxon Signed Rank test to determine whether there is a statistically significant difference between the unsupervised and supervised sets of weighting values (see Table 10). The following null hypothesis was therefore adopted: the median of differences between each pair of data sets (e.g., Supervised MD and Unsupervised MD) is equal to 0 (i.e., no statistically significant difference exists between the two).

Table 10. Tests to determine whether there is a statistically significant difference between an Unsupervised and a Supervised TLX weighting exercise (i.e., pairwise comparison)

Null Hypothesis[1]	*Significance (.025)*[2]	*Decision*
Supervised MD and Unsupervised MD	.000	Reject the NH
Supervised PD and Unsupervised PD	.000	Reject the NH
Supervised TD and Unsupervised TD	.304	*Retain* the NH
Supervised OP and Unsupervised OP	.021	Reject the NH
Supervised E and Unsupervised E	.011	Reject the NH
Supervised F and Unsupervised F	.000	Reject the NH

[1] **Null Hypothesis (NH):** The median of differences between each pair of data sets (e.g., Supervised MD and supervised MD) is equal to 0 (i.e., no statistically significant difference exists between the two)

[2] A comparison of two tests under different conditions is being presented using a Bonferroni adjusted alpha level (0.05/2 = 0.025)

Raw TLX vs mean weighted workload

Table 11 shows the medians for MWW and Raw TLX workload (RTLX) together with their respective deviations from the mean. RTLX does not take workload dimensions' weighting into consideration and is calculated by dividing the sum of all workload dimensions' raw ratings for each task/participant by six, the total number of dimensions. Eliminating this final pair-wise comparison to generate the MWW may in turn simplify the TLX process even further. To test this hypothesis a Spearman's rho correlation was run on the non-normally distributed values for MWW and RTLX. Two tests were carried out, one on the data collected during the follow-up workshops (117 observations from 13 participants reporting on nine fictitious tasks) and another test on values reported through the online questionnaire (94 students who had to enrol for an e-ID before using the e-service). In both cases the Spearman's rho revealed a positive and statistically significant relationship between MWW and RTLX (r_s [117] = .989, $p < .001$ and r_s [94] = .937, $p < .001$ respectively). In line with these observations, Cao et al. (2009) observed that RTLX

is more commonly adopted over MWW, citing the high correlation between weighted and unweighted workload scores as the main determining factor.

Table 11. This table shows the set of nine fictitious enrolment tasks together with their respective median MWW values alongside the median RTLX values

Task	*MWW*	*St. Dev.*	*RTLX*	*St. Dev*
A	0%	2.2	0%	2.6
B	0%	9.7	0%	8.1
C	18%	15.6	16%	13.3
D	11%	21.8	8%	17.7
E	32%	28	33%	22.3
F	14%	18.6	13%	17.1
G	12%	9.1	13%	8.8
H	21%	13.3	21%	13
I	81%	27.3	72%	24.4

Discussion

The use of NASA-TLX to measure perceived workload in the exam registration process and e-ID enrolment (where applicable), provided the author with very useful insights. This, together with data from follow-up sessions, helped to understand how students related to NASA-TLX's terminology and processes, as originally introduced by Hart and Staveland in (1988), with the aim to maximise NASA-TLX's validity and useability for this group of users and within this context.

Workload manifests itself in different ways

Students who have used the exam registration e-service, but had to go through the e-ID enrolment process, were expected to give significantly higher overall workload ratings than those who already had an e-ID, mainly due to the additional physical and temporal workload involved in travelling and queuing. However, this was not found to be the case, as there is a negligible difference in overall MWW between the two groups. By drilling down into NASA-TLX's multi-dimensional results it was noticed that sources of workload were significantly different for the two groups. Both presented a high measure of overall workload, albeit for different reasons. In principle those who had to enrol for an e-ID were concerned with delays and interruptions to their primary task; however, they indicated that the exam registration process was – in comparison – acceptable. On the other hand, students who already had an e-ID based their feedback mainly on the non-functional aspects of the exam registration process, such as lack of clarity in the process and site loading speed, resulting in a high level of frustration. Uni-dimensional workload measurement techniques do not explain the user experience in its entirety. Issues in design and performance can cause frustration, and this can be an equally important contributor to perceived workload, together with the more traditionally accepted sources of workload such as the physical and cognitive demands. The author recommends the adoption of a multi-dimensional workload assessment tool in order to understand the various sources of workload for different service alternatives. Future governments depend on the trust of younger citizens, and the interaction with

government institutions is formative for trust perceptions. Riegelsberger and Sasse (2010) point out that trust depends on the users' perception of motivation and competence – so being confronted with less than competently designed e-government services will undermine young people's trust in government.

Demystifying workload dimensions

Although provided with on-screen guidelines, participants in follow-up sessions were at times confused while rating certain dimensions, especially Own Performance (P), Effort (E) and Temporal Demand (TD). In particular Temporal Demand (TD) caused a level of confusion in its interpretation. Participants were often confused if Temporal Demand refers to how long it took to complete the task, or how long it should have taken.

Temporal Demand (TD) was originally introduced in NASA-TLX as a measure of time related pressure during a task, specifically the pace at which tasks occurred. This is a very context specific dimension especially suited for critical scenarios such as an emergency landing of an aircraft in bad weather. As is, this dimension may not be adequate for non-critical and mundane tasks. Further to this, some participants also voiced their concern on the similarity of certain workload dimensions. They explained:

> The main problem is that some of them are really similar. And you wouldn't know what to choose.

It was a non-trivial task to help participants understand the difference between the more abstract workload dimensions such as: Frustration (F) and Own Performance (P) or Effort (E) and Mental Demand (MD). Students were given the opportunity to think aloud and clarify their doubts throughout the exercise by asking questions. As one participant said:

> The only thing which struck me was the 'own performance' rating. Sometimes it is a bit hard to figure out what you did right or wrong so it's kind of hard to assess own performance.

Another comment related to how participants felt while conducting the final pairwise comparison, especially when they were asked to choose between Physical (PD) and Mental Demand (MD):

> Participant A: I also feel lazy with my choices.
>
> Participant B: True, true, same here.

In this case, both participants felt uncomfortable disclosing the fact that they preferred mental demand rather than physical demand; therefore, it can be seen that lack of anonymity may influence feedback. This ties in with Malheiros's (2014) observations on disclosure, whereby participants are less likely to disclose information comfortably and honestly if it portrays them in a bad light.

Keep out of reach of digital natives?

A series of tests, presented in Table 10 indicate that a supervised TLX exercise will yield a significantly different result in the way the six workload dimensions are weighted by digital natives. In the follow-up sessions the facilitator explained each and every workload dimension before going through the different tasks. This might have contributed towards the variance in interpretation, and thus in weighting outcomes, between online and workshop participants. Table 9 shows the differences in the interpretation of rating scales with and without supervision.

It was noticed that this group of users did not fully understand the official NASA-TLX descriptions for the various workload dimensions, in particular those for Mental Demand (MD), Effort (E) and Own Performance (P). Specific and age-appropriate examples were found to be helpful.

NASA-TLX, e-government enrolment and digital natives — does it really work?

Can this technique be used to measure workload confidently with digital natives? This section will tackle a subset of tasks from the nine fictitious enrolment processes presented during the follow-up sessions and their respective workload ratings across the six dimensions. Statistical tests show that there is a significant correlation between the resulting ratings and the demands imposed by the task. Figure 6 represents the overall mean adjusted ratings for three of these fictitious tasks, across the six workload dimensions. Service D had no major workload issues; however Temporal Demand (TD) and Frustration (F) were rated as being considerably high as the task required three days for account activation. Service G had low levels of workload across all dimensions; however, Mental Demand (MD) was the highest rated dimension. This can be explained by the fact that participants had to come up with a new password, a password hint and a call-in PIN to be used to authenticate themselves in case they need to call a help-desk. Service *I* had the highest ratings across all dimensions, and this was especially evident in Physical Demand (PD), Temporal Demand (TD), Effort (E) and Frustration (F). Half a day of travelling and queuing is required to complete the identity verification process as well as a three day period until the activation PIN is received by post.

Table 12. This table shows three different tasks from the set of nine fictitious enrolment tasks – denoting the participants' perceived mean weighted workload (MWW)

Task	*ItR*[1]	*ItG*[2]	*I*[3]	*D*[4]	*MWW*
D	4	2	No	Major[5]	11%
G	6	4	No	No	12%
I	NA	3	Yes[6]	Major[7]	81%

[1]Items to Recall [2]Items to Generate [3]Interruptions to daily routines [4]Delays [5]Wait three days before account is activated [6]Visit enrolment office during specific opening hours [7]Three day waiting period till an activation PIN is received by post

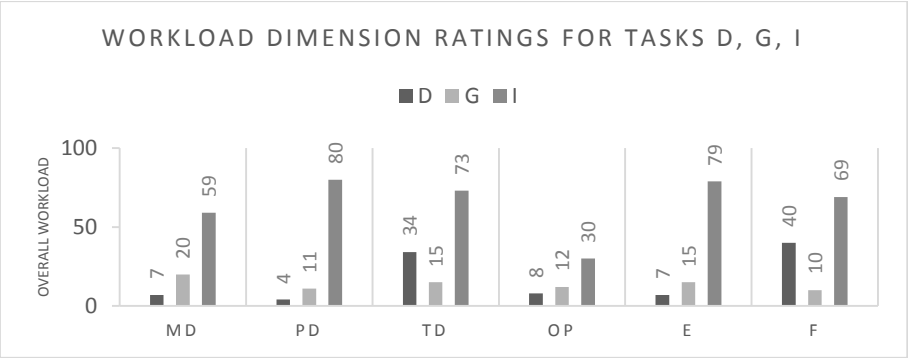

Figure 6. This chart shows the overall mean workload for the three tasks listed in Table 12

A degree of consistency was observed between the perceived workload for enrolment processes used on popular online services (see Table 3) and median weighted workload values for the nine enrolment tasks for fictitious services (see Table 7). Some noticeable examples are provided in Table 13. Although the two sets of results are close, one cannot exclude the possibility of other design factors placing significant influence on workload, especially on dimensions such as Frustration (F) and Effort (E).

Table 13. Contrasting perceived enrolment workload (PEW) derived by consensus from actual enrolment processes with TLX-based Mean Weighted Workload (MWW) values for similar, but fictitious tasks

Real Service	*PEW*	*Fictitious Service*	*MWW*
Hotmail	28%	Task H	21%
Gmail	13%	Task G	12%
National e-ID	81%	Task I	81%

Furthermore, following a series of tests presented in Tables 10 and 11, it was determined that even though the nine fictitious tasks were presented in a random order, on average participants reported statistically significant differences in perceived workload for tasks designed to be more demanding.

A final set of tests sheds more light on the need to retain the pairwise comparison exercise that is used to produce weighted workload values for each participant. Results provided in Table 11 show that there aren't any major advantages for the adoption of MWW values over RTLX, given the additional effort required from participants to complete the final pairwise comparison. Eliminating this final step may in turn simplify the TLX process even further.

Modifying NASA-TLX for use in e-service enrolment

The meaning of Temporal Demand (TD) may need to be modified to fit within an e-government context. The experience of 'feeling rushed' may not be an appropriate measure for enrolment processes, as opposed to other situations such as engaging

landing gears during an emergency landing. In the follow-up sessions Temporal Demand (TD) was expressed as a measure of the time required to complete the task. The associated hint should read: 'How much time did you require to complete this task?' This represents the perceived amount of time taken-up by the enrolment portion of an e-service, rather than the pressure exerted from time limitations.

Simpler definitions and context specific examples are needed for most of the rating scales:

- *Own Performance*: How confident were you during the enrolment process? Was the process easy to follow?
  The inverted labels for Own Performance (Good to Poor rather than Low to High) did not seem to be problematic.
- *Physical Demand*: How much physical effort did the process involve? Did you have to search for some documents? Did you need to go somewhere in-person to complete the transaction?
- *Mental Demand*: How much thought was required during this process? Did you have to come up with new secrets, such as usernames, passwords or PINs? Did you have to provide a lot of information to complete the form(s)?
- *Effort*: Considering both mental and physical demand, did it require a lot of effort to perform the process?
- *Frustration*: How irritating or annoying was this enrolment process?

If possible provide a channel for immediate feedback during the TLX rating process using voice over IP (VoIP) if physical proximity is not possible. Finally, Raw TLX was found to be a suitable measure to inform designers about the perceived workload for this group of users (digital natives), while also simplifying the overall rating process. This was mainly due to the fact that a high level of correlation was found between Raw TLX and MWW values, making the additional effort required to generate MWW values unjustifiable.

Conclusions

Following a rigorous empirical exercise, this paper offers insights on the applicability of NASA-TLX as a highly-cited human factors technique to measure the impact of enrolment process design on e-government service users. The literature reviewed positions NASA-TLX as one of the better workload assessment techniques, in both sensitiveness and ease of use. It has been adopted in a number of domains and applications, from analysing flight crew complement requirements and down to evaluating software interfaces. This study's goal was to shed more light on the effectiveness of NASA-TLX, particulary when used by and on digital natives in an e-government context.

NASA-TLX provided interesting insights into the possible sources of workload for this group of users, and it was found to be fairly sensitive to changes in workload parameters, informing the researcher of possible actions to reduce workload perceptions, improve adoption and if compulsion exists, minimise resentment. With minor modifications NASA-TLX could be improved to serve its purpose better

within this particular context and with this user group. This also includes additional guidance on the meaning and implications of the various workload dimensions. Finally, it has been noted that in this context there are no major advantages arising from the use of the MWW metric over RTLX.

References

Axelsson, K. & Melin, U. (2012). Citizens' attitudes towards electronic identification in a public e-service context an essential perspective in the eid development process. In Hans J. Scholl, Marijn Janssen, Maria A. Wimmer, Carl Erik Moe, and Leif Skiftenes Flak, editors, *Electronic Government*, volume 7443 of Lecture Notes in Computer Science, pp.260–272. Springer Berlin Heidelberg.

Cain, B. (2007). *A review of the mental workload literature*. Technical report, Toronto, Canada: Defence Research and Development.

Cao, A., Chintamani, K.K., Pandya, A.K., & Ellis, R.D. (2009). NASA-TLX: Software for assessing subjective mental workload. *Behavior Research Methods*, 41(1):113–117.

Card, S.K., Moran, T.P., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Commun.ACM*, 23(7), 396–410.

GARTEUR. (2003). Action Group FM AG13. *Garteur handbook of mental workload measurement*. Technical report.

Hart, S. G. (2006). Nasa-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*, 904–908.

Hart S.G. & Staveland L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology - Human Mental Workload*, *52*,139– 183.

Hill S.G., Iavecchia H.P., Byers J.C., Bittner A.C., Zaklad, A.L., & Christ, R E. (1992). Comparison of four subjective workload rating scales. *Human Factors, 34*, 429-439.

Inglesant, P. & Sasse, M.A. (2007). Usability is the best policy: Public policy and the lived experience of transport systems in London. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not as We Know It* - Volume 1, BCS-HCI '07, pp. 35–44. British Computer Society.

Katsanos, C., Karousos, N., Tselios, N., Xenos, M., & Avouris, N. (2013). KLM form analyzer: Automated evaluation of web form filling tasks using human performance models. In *Human-Computer Interaction - INTERACT 2013*, volume 8118 of Lecture Notes in Computer Science, pp. 530–537. Springer Berlin Heidelberg.

Malheiros, M. (2014). User behaviour in personal data disclosure. PhD thesis, University College London.

Porter, C., Sasse, M. A., & Letier, E. (2012). Designing acceptable user registration processes for e-services. In Proceedings of *HCI 2012 - The 26th BCS Conference on Human Computer Interaction*. BCS.

Porter, C., Sasse, M. A., & Letier, E. (2013). Giving a voice to personas in the design of e-government identity processes. In Proceedings of *HCI 2013 - Research*

to Design: Challenges of Qualitative Data Representation and Interpretation in HCI. BCS.

Prensky, M. (2001). Digital natives, digital immigrants. *On the horizon*, *9*(5), 1–6.

Riegelsberger, J. & Sasse, M. A. (2010). Ignore these at your peril: Ten principles for trust design. In *Trust 2010. 3rd International Conference on Trust and Trustworthy Computing*.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of swat, NASA-TLX, and workload profile methods. *Applied Psychology*, *53*, 61–86.

Sasse, M. A., Steves, M., Krol, K., & Chisnell, D. (2014). The great authentication fatigue - and how to overcome it. In P.L.Patrick Rau, editor, *Cross-Cultural Design*, volume 8528 of Lecture Notes in Computer Science, pp 228–239. Springer International Publishing.

Steves, M., Chisnell, D., Sasse, M. A., Krol, K., Theofanos, M., & Wald H. (2014). Report: Authentication diary study. NISTIR 7983. *Technical Report NISTIR 7983*.