Investigating the Task Load of Investigating the Task Load in Visualization Studies

Daniel Pahr[†] and Sara Di Bartolomeo[‡]

[†]University of Vienna and [‡]TU Wien, Vienna, Austria.

This document was made to resemble the original NASA TLX Paper and Pencil Package (https://humansystems.arc.nasa.gov/groups/tlx/downloads/TLX_pappen_manual.pdf).

Abstract

The NASA task load index (short: NASA-TLX) is a common metric to evaluate the workload of a user in a visualization study. Yet, it is rarely performed as initially intended, as the sources-of-workload evaluation is often omitted for various reasons. We conduct an online survey to investigate the task load of administering different versions of the NASA-TLX in a meta-study using the ReVISit framework. Our results show that it is not the slight increase in experiment time, but rather participants' frustration with the procedure, that contributes to the slight increase in task load when using the full version of the TLX compared to using a shortened version. However, we also show that the full version can shine a different and more faceted light on workload by adding a personal dimension to the data. We propose that a compact version of the sources-of-workload questionnaire can mitigate both time loss and frustration for study participants, while still providing the same data as the original procedure. The online study can be found and interactively explored on https://dpahr.github.io/tlxtlx/**, and the source for the study, as well as the code for our analysis, can be found on** https://github.com/dpahr/tlxtlx/**.**

1 Introduction

A central element of many evaluations — in the field of visualization and beyond — is the measurement of the subjective task load on a user. This can, for example, shine a light on the potential of a tool to

lighten the cognitive load of users in comparison to other state-of-the-art methods. Among all the tools in the human-factors toolbox, few have earned the kind of celebrity status that the NASA Task Load Index (NASA-TLX) enjoys. Born in the cockpit but well established in usability labs, hospitals, control rooms, and even classrooms, NASA-TLX has become the de facto method for asking people one simple thing: How hard was that task, really? This is done by rating six separate items on a questionnaire, dividing task load into its individual parts: mental, physical, and temporal demand, performance, effort, and frustration.

The procedure for the NASA-TLX is well illustrated and disseminated via online resources, such as the paper and pencil package [4] or the computerized version [3]. Nevertheless, the procedure is often modified [5], making it difficult to compare results between different publications and distorting the original meaning of this important metric. The most common modification is to omit performing a user-specific sources of workload evaluation, which provides weights for the individual dimensions of the TLX.

We argue that, specifically for the analysis of visualizations using modern survey platforms, the complete procedure for the NASA-TLX is not only easy to implement but also quickly executed. Thus, in this paper, we turn the tables. Instead of using NASA-TLX to evaluate a tool, we evaluate NASA-TLX itself. Our goal is to measure the impact of applying different versions of the NASA-TLX in a crowdsourced online study. We create a survey using the reVISit framework [2], that confronts participants alternatingly with i) no scale weighting, ii) a compact scale rating procedure, and iii) a scale rating procedure that resembles the original paper version.

By measuring the impact of using the full version of the NASA-TLX, we demonstrate that this version offers more insights with a reasonable impact on study participants. Furthermore, we argue that this impact can be mitigated by using a compact version of the sources-of-workload evaluation.

2 Background

The NASA-TLX, developed by Hart and Staveland [6] in 1988, breaks subjective workload down into six intuitive components: mental demand, physical demand, temporal demand, effort, performance, and frustration. This can be done by marking a rating on a piece of paper or a PC, on a questionnaire similar to Figure 1. A slow interface might lead to user frustration, while a complicated task may lead to a higher cognitive load. As task load is a highly subjective measure, each dimension is given a different weight for every individual, which can also vary across different tasks. For example, one might care very little about temporal demand when there is no time limit for a task, or disregard physical effort when simply interacting with a computer using a mouse and keyboard. The TLX is designed to account for participant-specific sources of workload. For each type of task, a participant is presented with all 15 possible pairings of the six scales in a random order, and for each, selects the one contributing higher to their subjective task load. The participant then rates their experience in each of the categories on a scale from 0 to 100, and the final score — the task load index — is computed as a weighted average of the individual scores.

However, the procedure to acquire user-specific weights for the individual components is often disre-

MENTAL DEMAND

How mentally demanding was the task?

Very Low | Very High

PHYSICAL DEMAND

How physically demanding was the task?

Very Low | Very High

TEMPORAL DEMAND

How hurried or rushed was the pace of the task?

Very Low | Very High

PERFORMANCE

How successful were you in accomplishing what you were asked to do?

Good | Bad

EFFORT

How hard did you have to work to accomplish your level of performance?

Very Low | Very High

FRUSTRATION

How insecure, discouraged, irritated, stressed, and annoyed were you?

Very Low | Very High

Figure 1: A NASA-TLX questionnaire, as proposed by the paper and pencil package [4].

garded. This has sometimes been justified through a need for efficiency, since the weighting procedure takes time [9]. Some research has been done to prove it to be redundant, due to a high correlation of the weighted and non-weighted TLX in some cases [1, 7, 9]. The abridged version without collecting individual weights for the scales has been called the "raw" task load index (RTLX). Hendy et al. [7] suggest that even a univariate rating, i.e., using a single scale for task load, would provide a good estimate of individual workload. We argue that using the complete procedure over the RTLX can provide additional insights that are otherwise lost. While raw ratings may be used to pinpoint specific points of interest in an analysis, a raw average may vary greatly from the weighted average in specific cases [6].

Multiple versions of the TLX questionnaire are available directly from NASA, with precise instructions on how to use them. The paper and pencil package [4] provides printable resources for a physical version, a computerized version instead proposes filling the questionnaires on a PC [3], even an iOS version is provided [12]. Noyes and Bruneau [10] compare the original paper and pencil package and the computerized version in a meta study, investigating if either of these versions incurs a higher task load on their study participants. They found the two versions to incur comparable task load. Our study, while also a TLX metastudy where the TLX is applied as a metric to evaluate itself, is targeted to investigate different versions of the virtual questionnaire.

Kosch et al. [8] warn of the "hidden cost of the NASA-TLX". They argue that the TLX was not developed with HCI in mind, and that a cumulated score may obscure individual factors contributing to workload. Still, they argue that its simplicity and ease of use are key advantages over other metrics. In this paper, we seek to disentangle simplicity and malpractice using the NASA-TLX. We argue that using the RTLX brings no significant advantage over performing the full procedure. The slightly longer procedure can be administered with little impact on study times and user experience, using modern survey platforms.

3 Experimental Setup

We implemented a small survey in reVISit [2], comparing the impact of three different versions of the NASA-TLX on the task load of a study participant in a between-subjects design. The online study can be found and interactively explored on our ReVISit instance, and the source for the study, as well as the code for our analysis, can be found on the GitHub repository.

Procedure

We show an illustration of the experiment procedure in Figure 2. Our study starts with an introduction for every participant. We give a brief overview of the procedure and make sure that participants are aware that they will complete a two-part experiment. The first part comprises the participant performing a single visualization task from the mini-VLAT questionnaire [11] and being administered one of three versions of the NASA-TLX. This first execution of the TLX is targeted to evaluate the participants' experience in completing the task.

The second part has the participant completing another NASA-TLX questionnaire. This is always the full procedure, including the sources-of-workload evaluation, with the scales presented in pairs in a random order. We carefully introduce the second part

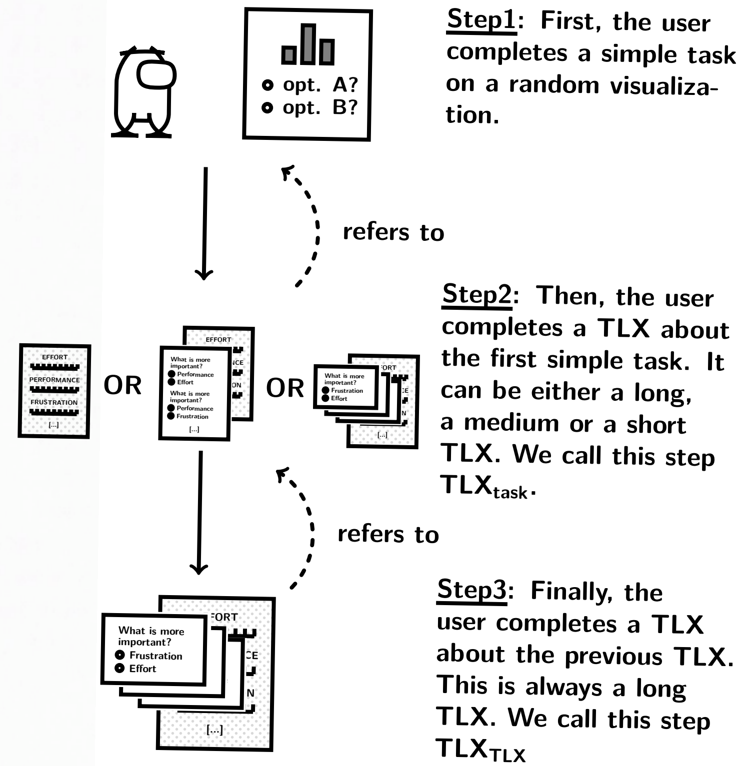


Figure 2: Illustration of the experiment procedure in 3 steps.

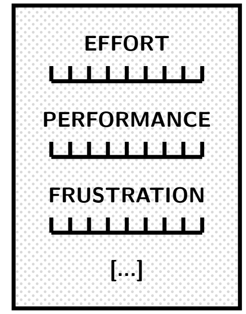
with a title card and a thorough introduction to make sure subjects understand that they are now evaluating the TLX procedure from part one.

Conditions

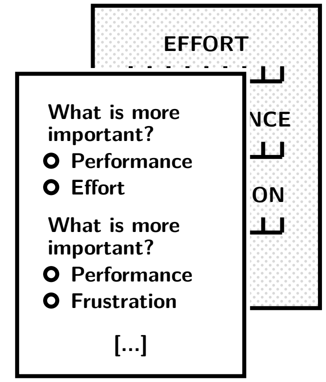
We compare three different versions of the NASA-TLX in our study (*long, medium, short*). Figure 3 shows an overview of our experiment conditions. The *long* condition represents the full TLX procedure, most closely resembling the original paper and pencil package. Participants are first shown pairs of scale names in a random order, selecting the one that they deem to contribute more to their subjective task load. Afterwards, they are asked to rank their task load for the specific task on the six scales. The task load index is calculated as a weighted average of a user's ratings using the results of the sources-of-workload procedure. As a *medium* condition, with a predicted effort between the RTLX and the original version of the TLX, we use a compact form of the sources-of-workload evaluation. Instead of presenting the pairs of scales in succession, participants are shown a single page with all 15 pairs at once. The task load is computed in the same way as in the *long* condition. In the *small* condition, the NASA-TLX is administered in its abridged form, commonly known as the RTLX. The difference to the original version is that no sources-of-workload evaluation is performed, thus no weights are obtained for the individual scales when computing the TLX score. The task load is computed as the average of a user's ratings.

Metrics

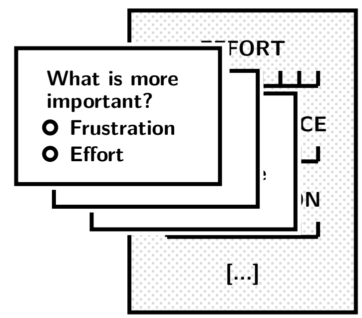
Our goal is to determine the impact of using the NASA-TLX on participants in a crowdsourced study.



(a) Condition (short): only the sliders are presented.



(b) Condition (medium): Sources-of-workload on a single page.



(c) Condition (long): sources of workload on individual pages.

Figure 3: The three different conditions for the experiment. (a) Participants only complete the TLX questionnaire, no sources-of-worload evaluation (short). (b) Participants complete the sources-of-workload evaluation on a single page (medium). (c) Participants complete the sources-of-workload evaluation with successive comparisons, analogous to the paper and pencil package (long).

The task the participants are asked to perform initially serves as a source of exertion, i.e., to create the workload we wish to evaluate. Our participants complete two separate NASA-TLX questionnaires, one to determine the task load of the visualization task (TLX_{task}) and the other to determine the task load of the first administered TLX questionnaire (TLX_{TLX}). In the *long* and *medium* conditions, we look at the correlation between the weighted (TLX_{task}) and unweighted ($RTLX_{task}$) average of the ratings for the visualization task. Additionally, we compare the results from the sources-of-workload evaluation ($w_{TLX_{task}}$) and the participants' questionnaire responses ($RTLX_{task}$) in rating the visualization task. For the evaluation of the TLX procedure, we look at time ($t_{TLX_{task}}$) and task load index (TLX_{TLX}) of our participants across the three conditions (*long, medium, short*). The time for completing the questionnaire is measured from the participant having read the introduction, which either leads to the sources-of-workload evaluation or directly to the rating questionnaire, to the completion of the rating questionnaire. We compute the TLX for a task as a weighted average of scores and weights per scale, and the RTLX as the average of scores on each scale.

4 Results

We sent out an invitation to participate in our study via email to several visualization research groups. We offered no reward for participation and did not collect demographic data. Of 34 responses in total, 20 completed the entire study, and we rejected 14 incomplete responses. The number of samples received for each condition varied; we received seven responses for the *long*, three responses for the *medium*, and 10

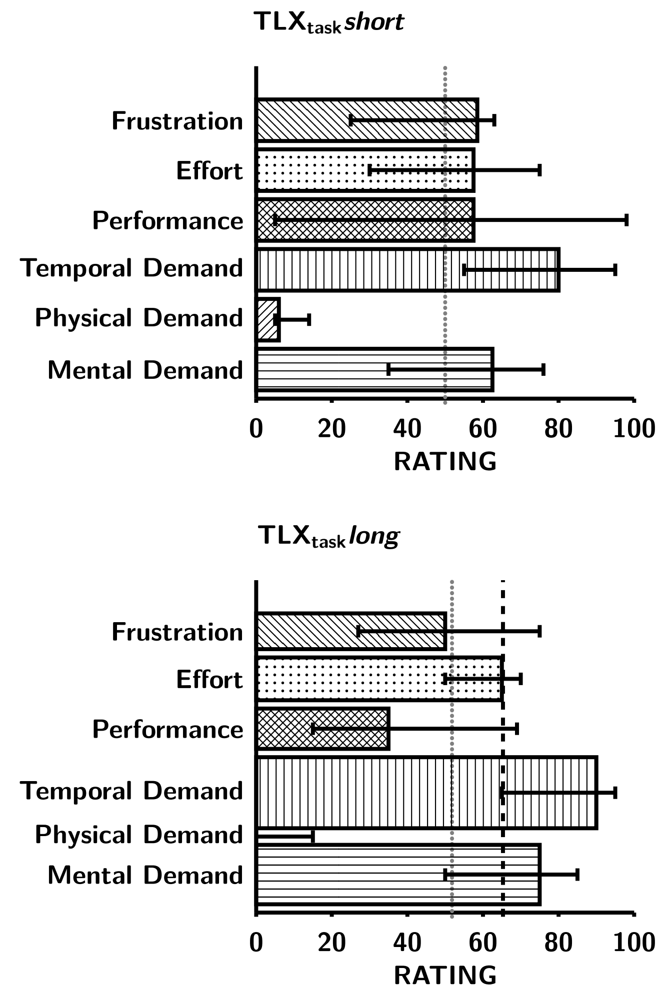


Figure 4: The results for TLX_{task}. A black dashed line indicates the average TLX score, while the grey dashed line indicates the unweighted average RTLX score. As we did not collect weights in the *short* condition, only the RTLX score is shown.

responses for the *short* version. We focus mainly on the comparison of the *long* and *short* versions, and only briefly discuss implications of our findings on the *medium* version due to the low number of responses in this condition.

We present the results similar to Hart [4] in their paper and pencil package. They use a two-dimensional representation, in the form of a bar chart, where the length of the bars represents the rating for the individual scale and the width of the bar represents the chosen weight. We use this representation to show the median rating per scale and condition, and whiskers representing the quartiles. For the *long* version, we display the average importance rating for each scale as the width of the bar. The overall average workload, i.e., the TLX, is shown as a dotted line across the chart.

Evaluation of a Visualization Task

Figure 4 shows the results of the task load evaluation for the mini-VLAT task that our participants completed.

Comparing the RTLX of the *long* version, i.e., calculating the raw average of scores instead of the weighted average, we see the average task load index roughly equal to the *short* condition (51.81 vs 50). The RTLX is represented by the grey lines in Figure 4, while the black lines represent the weighted TLX score. Looking at the weighted average in the *long* condition shows that the overall workload in the *long* condition now much higher than in the *short* condition (65.29 vs 50), while the ratings on the scales themselves are very similar. We see an explanation for this in the average weighting. Temporal and mental demand are weighed higher on average than frustration, effort, and performance, thus increasing

the overall task load. On the other hand, physical demand was consistently rated unimportant by the participants, and does not have an impact on the TLX in the *long* version, while in the *short* version, it brings down the overall score.

Evaluation of a TLX Questionnaire

Figure 5 shows the results of the task load evaluation for the task load index questionnaire our participants completed.

The median completion time for the TLX evaluation of the task was 81 seconds in the *short* condition, which is almost twice as fast as in the *long* condition with 156 seconds. For the average task load, we see an increase in the *long* version compared to the *short* version (52.43 vs 43.10). Interestingly, while we did measure that participants spent almost double the amount of time on the *long* version than on the *short* version (156 seconds vs 81 seconds), temporal demand in completing the task load questionnaire did not seem to impact our participants' task load much in either condition. The most apparent difference between the two conditions is the increased median frustration of the participants in the *long* version compared to the *short* version (70 vs 40), possibly caused by the high number of pairwise comparisons presented in succession.

The *medium* Version

We only collected three samples for the *medium* condition, hence we omit discussing the results of our analysis in comparison with the other two conditions. However, we still see it worthwhile to discuss the initial results in this separate section.

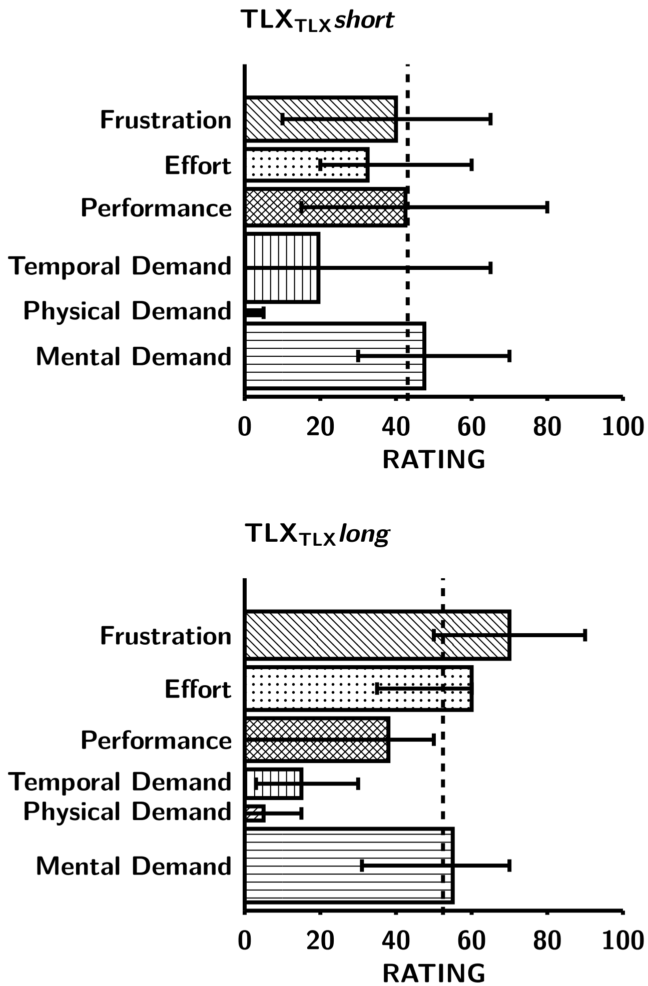


Figure 5: Results for TLX$_{TLX}$. A black dashed line indicates the average TLX score.

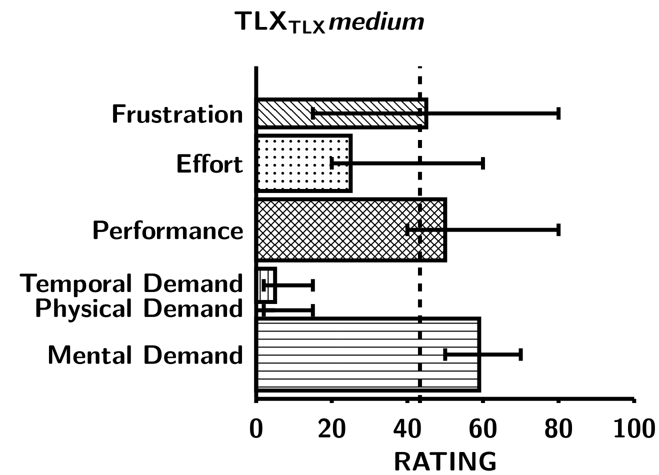


Figure 6: Results from TLX$_{TLX}$ for the *medium* condition. A black dashed line indicates the average TLX score.

Figure 6 shows the results for the experiment from the medium condition. The average overall workload for the *medium* condition is similar to the *short* condition (43.33 vs 43.10). We observe again reduced temporal demand reported for this condition, even compared to the other two. The median rating for frustration in the *medium* condition is also less than in the *long* condition (45 vs 70), however the highest reported frustration score was 80 in the *medium* condition. The average completion time of the *medium* questionnaire was 118 seconds, which is faster than the average (252 seconds) and median (156 seconds) of the *long* version. While these results are promising for an initial observation, they should still be taken with a grain of salt, due to the low sample count.

Discussion

At first sight, our study results seem to confirm the validity of criticism on the length of the full TLX procedure by Moroney et al.[9]. Our participants who completed a sources-of-workload evaluation spent more time than those who did not, leading to a longer total time participants spent on the experiment. However, our procedure saw participants complete only a single task in our experiment. The sources-of-workload procedure is only completed once per task type; hence, for longer experiments, the additional time spent on the sources-of-workload evaluation also becomes relatively less of a contributor to the experiment time.

In their 20-year anniversary review [5] of the NASA-TLX, Sandra G. Hart, one of the researchers behind the TLX, reports:

> In the 29 studies in which RTLX was compared to the original version, it was found to be either more sensitive (Hendy, Hamilton, & Landry, 1993), less sensitive (Liu & Wickens, 1994), or equally sensitive (Byers, Bittner, Hill, 1989), so it seems you can take your pick.

We argue that the sources-of-workload procedure effectively measures a second dimension of task load, which we try to indicate through encoding the average weights per scale into our figures. Our representation shows personal preferences for a task by revealing the sources of workload of a study sample.

While we did not receive enough responses, we still observe an improvement in completion time and task when using the compact sources-of-workload evaluation compared to the original version. An advantage of the single-page version of the sources-of-

workload evaluation may be that study participants can see how many pairs there are for them to rate, mitigating their frustration.

We received little written feedback in our survey, the only comment stating that the participant got confused with the scale names. While we provided the descriptions as help text in ReVISit, we argue that users need consistent support in completing the TLX, especially in crowd-sourced studies. Another indicator for this is the high variance of the performance score in most of our experiments, which is counterintuitive in the TLX questionnaire since it ranges from “good” to “poor”, while other scores all range from “low” to “high”. Here, we suggest including scale definitions directly on screen whenever referring to them, in any virtual setting.

Frequently, the TLX questionnaire is modified even beyond omitting the sources of workload evaluation. In visualization studies, commonly performed in front of computer screens, using mainly mouse and keyboard, the physical component has been called into question or outright disregarded [13]. This may be sensible in isolated A/B comparisons; however, by modifying an established metric, the ability to compare results with prior studies is lost. Beyond this, opting to report on individual scales instead of a compound score complicates statistical analysis, requiring adjustments in hypothesis testing such as Bonferroni correction. The TLX is designed to be a metric for task load; hence, if one desires to measure specific sub-items, then assumptions can only be made for that specific item, not for task load overall, and the reference to the TLX would be superfluous.

Conclusion

Kosch et al. [8] speak of the "hidden cost" of the TLX. Instead, we want to highlight its forgotten benefits. Firstly, the NASA-TLX, while not initially intended for use in HCI, measures highly relevant dimensions of workload independent of the nature of the task. Secondly, the personalized nature of the questionnaire allows us to analyze an otherwise hidden dimension of workload, lost to us if the sources-of-workload evaluation is not performed. Finally, modern study frameworks allow us to execute a procedure that may have been cumbersome in the past with considerable ease. The NASA-TLX remains a useful measurement tool in our studies, now even for itself.

Acknowledgement

The authors thank Tingying He from the University of Utah for providing us with the implementation of the NASA-TLX in ReVISit.

Conflict of Interest Statement

Sara Di Bartolomeo is also involved in the organization of the workshop this document is submitted to, alt.vis 2025.

About the Look of this Document

The look of the document is meant to emulate the appearance of the original NASA TLX Pen and Paper manual. What you are looking at is not a messy scan of a document — instead, everything about this document is completely generated with LaTeX, including

scanlines and stapling in the middle of the booklet. All of the charts, drawings and diagrams are also just commands, generating everything using tikz. The source code for generating this document in LaTeXwill be distributed as part of the supplemental material.

References

[1] James C. Byers, A. C. Bittner, and Susan G. Hill. "Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary". In: *Advances in industrial ergonomics and safety* 1 (1989), pp. 481–485.

[2] Yiren Ding, Jack Wilburn, Hilson Shrestha, Akim Ndlovu, Kiran Gadhave, Carolina Nobre, Alexander Lex, and Lane Harrison. "reVISit: Supporting Scalable Evaluation of Interactive Visualizations". In: *2023 IEEE Visualization and Visual Analytics (VIS)*. Oct. 2023, pp. 31–35. DOI: 10.1109/VIS54172.2023.00015.

[3] Sandra G. Hart. *NASA Task Load Index (TLX): Computerized Version - Volume 1.0*. Jan. 1986.

[4] Sandra G. Hart. *NASA Task Load Index (TLX): Paper and Pencil Package - Volume 1.0*. Jan. 1986.

[5] Sandra G. Hart. "Nasa-Task Load Index (NASA-TLX); 20 Years Later". EN. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50.9 (Oct. 2006), pp. 904–908. DOI: 10.1177/154193120605000909.

[6] Sandra G. Hart and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: *Advances in Psychology*. Ed. by Peter A. Hancock and Najmedin Meshkati. Vol. 52. Human Mental Workload. North-Holland, Jan. 1988, pp. 139–183. DOI: 10.1016/S0166-4115(08)62386-9.

[7] Keith C. Hendy, Kevin M. Hamilton, and Lois N. Landry. "Measuring Subjective Workload: When Is One Scale Better Than Many?" EN. In: *Human Factors* 35.4 (Dec. 1993), pp. 579–601. DOI: 10.1177/001872089303500401.

[8] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. "A Survey on Measuring Cognitive Workload in Human-Computer Interaction". In: *ACM Comput. Surv.* 55.13s (July 2023), 283:1–283:39. DOI: 10.1145/3582272.

[9] W.F. Moroney, D.W. Biers, F.T. Eggemeier, and J.A. Mitchell. "A comparison of two scoring procedures with the NASA task load index in a simulated flight task". In: *Proceedings of the IEEE 1992 National Aerospace and Electronics Conference@m_NAECON 1992*. May 1992, 734–740 vol.2. DOI: 10.1109/NAECON.1992.220513.

[10] Jan M. Noyes and Daniel P. J. Bruneau. "A self-analysis of the NASA-TLX workload measure". In: *Ergonomics* 50.4 (Apr. 2007), pp. 514–519. DOI: 10.1080/00140130701235232.

[11] Saugat Pandey and Alvitta Ottley. "Mini-VLAT: A Short and Effective Measure of Visualization Literacy". en. In: *Computer Graphics Forum* 42.3 (2023), pp. 1–11. DOI: 10.1111/cgf.14809.

[12] *TLX @ NASA Ames - NASA TLX App.*

[13] Eliane Zambon Victorelli, Anne-Flore Cabouat, Emanuele Santos, Florent Cabric, and Petra Isenberg. "We Should Change How We Measure User Experience in Visual Analytics Systems ". In: *EuroVis Workshop on Visual Analytics (EuroVA)*. Ed. by Hans-Jörg Schulz and Anna Villanova. The Eurographics Association, 2025. DOI: 10.2312/eurova.20251095.