Development and validation of a behavioral video coding scheme for detecting mental workload in manual assembly.

Bram B. Van Acker[abd], Davy D. Parmentier[a], Peter D. Conradie[ad], Stephanie Van Hove[cd], Alessandro Biondi[a], Klaas Bombeke[cd], Peter Vlerick[b], Jelle Saldien[ad]

Accepted author version

[a] Department of Industrial Systems and Product Design, Faculty of Engineering and Architecture, Ghent University, Technologiepark Zwijnaarde 46, 9052, Zwijnaarde, Belgium
[b] Department of Work, Organisation and Society, Faculty of Psychology and Educational Sciences, Ghent University, Henri Dunantlaan 2, 9000, Ghent, Belgium
[c] Department of Communication Sciences, Faculty of Political and Social Sciences, Universiteitstraat 8, 9000 Gent
[d] Research group IMEC-MICT-Ghent University, De Krook, Miriam Makebaplein 1, 9000 Ghent, Belgium

* Corresponding author: bramb.vanacker@ugent.be, +32 478 57 14 11, orcid.org/0000-0002-6565-3569

Word count = 9050 words

(excluding title, abstract, practioner summary, tables, figures, acknowledgments, references and appendices)

To cite this article, please refer to version of record with the citation:
Bram B. Van Acker, Davy D. Parmentier, Peter D. Conradie, Stephanie Van Hove, Alessandro Biondi, Klaas Bombeke, Peter Vlerick & Jelle Saldien (2021) Development and validation of a behavioural video coding scheme for detecting mental workload in manual assembly, Ergonomics, 64:1, 78-102.
DOI: 10.1080/00140139.2020.1811400

Abstract: Manual assembly in the future *Industry 4.0* workplace will put high demands on operators' cognitive processing. The development of mental workload (MWL) measures therefore looms large. Physiological gauges such as electroencephalography (EEG) show promising possibilities, but still lack sufficient reliability when applied in the field. This study presents an alternative measure with a substantial ecological validity. First, we developed a behavioral video coding scheme identifying 11 assembly behaviors potentially revealing MWL being too high. Subsequently, we explored its validity by analyzing videos of 24 participants performing a high and a low complexity assembly. Results showed that five of the behaviors identified, such as freezing and the amount of part rotations, significantly differed in occurrence and/or duration between the two conditions. The study hereby proposes a novel and naturalistic method that could help practitioners to map and redesign critical assembly phases, and researchers to enrich validation of MWL-measures through measurement triangulation.

Keywords: mental workload, measurement, behavioral video coding, validation, assembly

Practitioners summary: Current physiological mental workload (MWL) measures still lack sufficient reliability when applied in the field. Therefore, we identified several observable assembly behaviors that could reveal MWL being too high. The results propose a method to map MWL by observing specific assembly behaviors such as freezing and rotating parts.

1. Introduction

The current growing demand for mass-customization yields a greater product variety (Um, Lyons, Lam, Cheng, & Dominguez-Pery, 2017; Wan & Sanders, 2017) and thereby a higher manufacturing complexity (ElMaraghy, ElMaraghy, Tomiyama, & Monostori, 2012; Hu, Zhu, Wang, & Koren, 2008). Despite the extensive automation of production processes, manual assembly still remains highly valuable for the extreme flexibility it provides (Booker, Swift, & Brown, 2005). Cognitive demands on operators in this context and for sure in the future Industry 4.0 setup however increase accordingly, reinforcing the long-standing need for mental workload (MWL) measurement in order to optimize cognitive ergonomics (Van Acker, Parmentier, Vlerick, & Saldien, 2018; Young, Brookhuis, Wickens, & Hancock, 2014). With a thorough and objective understanding of non-optimal MWL (i.e., overload *or* underload) the presentation of information and materials can be redesigned (Brolin, Thorvald, & Case, 2017), assemblies can be made more intuitive (Parmentier, Van Acker, Detand, & Saldien, 2019) or smart technologies can be introduced to assist the operator (Erol, Jäger, Hold, Ott, & Sihn, 2016; Longo, Nicoletti, & Padovano, 2017).

Subjective measures have long been the preferred method to indirectly infer MWL (cf., Young et al., 2014), but are intrusive when applied in the field and only provide a subjective estimation (for a discussion, see de Waard & Lewis-Evans, 2014; de Winter, 2014; Matthews, De Winter, & Hancock, 2019) of accumulated load instead of more sensitive fluctuations in MWL-levels (Antonenko, Paas, Grabner, & van Gog, 2010). More recent methods offer a more direct and continuous way of measuring MWL through physiological reactions within the operator and this with a very limited latency (Charles & Nixon, 2019). EEG, for instance, or functional Near Infrared Spectroscopy (fNIRS) measuring frontal cerebral blood flow velocity have demonstrated exciting potential for future unobtrusive MWL measurement (cf., Foy & Chapman, 2018; Guru et al., 2015; Hairston et al., 2014; McKendrick et al., 2016) and even mental overload measurement (Morton et al., 2019). Such measures are nevertheless still in full development and do not yet achieve adequate effect sizes (Vanneste et al., 2020) nor adequate reliability or external validity in applied mobile settings, this due to numerous confounds such as movement and breathing artefacts (see Arico, Borghini, Di Flumeri, Sciaraffa, & Babiloni,

2018; Brouwer, Zander, van Erp, Korteling, & Bronkhorst, 2015). Additionally, differences in MWL are often estimated by comparing discrete conditions, instead of measurement being instantaneous. To arrive at the stage of measuring MWL fluctuations within a short time span - say, seconds - and in a real-world context, great strides need to be made yet.

Where the field of subjective and physiological MWL measurement validation assesses MWL changes to eventually detect or predict decrements in performance (see, e.g., Young et al., 2014), here, we reversed the logic, in line with primary task performance measures (cf., Cain, 2007). We designed a coding scheme by narrowly defining observable types of inefficient and ineffective assembly behaviors that potentially reveal MWL being too high. The current study is hence, to our knowledge, the first to explore behavioral observation as a measure of MWL in an ecologically valid assembly context. Specifically, it builds on current knowledge on human cognition in assembly performance (but not providing a validated measure yet) and translates this knowledge and similar work on behavior analysis in human-computer interaction and driving into a novel and validated mental workload measure for manual assembly. It hereby aims to help design a method already meeting some of the widely envisioned criteria for in-the-field MWL-measurement on the short term, i.e., non-obtrusiveness, sensitivity to rapidly changing MWL fluctuations and ease-of-use (see Matthews, Reinerman-Jones, Barber, & Abich, 2015).

2. Background

Scrutiny on the concept of MWL upholds a long tradition and even thrives during recent years, driven by the expectations for the upcoming Industry 4.0 (Young et al., 2014). Altogether, MWL exists as a function of task demands and moderating variables, and can be understood as a subjective experience and a physiological reaction, resulting in task-related behavior (Van Acker et al., 2018). Research on the measurement of MWL in industrial contexts has mainly focused on subjective estimations and physiological reactions (Charles & Nixon, 2019). Gauging the third, behavioral component in these settings has however rarely made the step beyond measuring, e.g., execution times, reaction times or errors. The same theorization explaining fluctuations in the subjective and physiological components of MWL can however also help explain fluctuations in observable behavior.

2.1 MWL revealing limited resources

MWL can be defined as a subjectively experienced physiological processing state resulting from the interplay between the human cognitive architecture and the work demands being attended to (Van Acker et al., 2018). Seminal work learns that the cognitive physiological resources propelling this interplay are limited (e.g., because of competition between visual and auditory sensory modalities; Wickens, 2002, 2008) and draw from a common and also limited underlying pool of physiological resources from which also competing emotional load (for a distinction between cognitive and emotional load, see Van Acker et al., 2018) and physical load draw (Kahneman, 1973; Mandler, 1979; Norman & Bobrow, 1975; for an overview, see Staal, 2004) (see Figure 2).

2.2 Performance revealing MWL

In assembly specifically, cognitive resources are allocated for perception, response selection and action, propelled by, i.a., attention and memory (Stork & Schubö, 2010). For commissioning, the perception stage include, i.a., identifying relevant parts. For joining, the operator's perception entails, for instance, processing the required part position and orientation. An appropriate response is then selected (e.g., 'take this part and join it on that position in that orientation') to finally go through the process of motor execution of the joining operation (including planning and adjusting). Since resources are limited, this concrete assembly performance can become less efficient and less effective when MWL increases. Commissioning, for example, can take longer, while response selection can entail taking two wrong parts first.

Previous research shows that the relationship between (assembly) performance and MWL exhibits an inverted U-shaped trend (de Waard, 1996; Montani, Vandenberghe, Khedhaouria, & Courcy, 2020; Young et al., 2014). In Figure 1 (inspired on de Waard, 1996; Hart & Wickens, 2008; Young et al., 2014), it can be observed that, in this way, performance is optimal when MWL resides around moderate levels (cf., flow theory; Bruya, 2010), while it is non-optimal (i.e., too low) in case of MWL being too low or too high. The MWL 'redlines' represent the performance break points here, so that at the

right side of the right redline, *overload*, and at the left side of the left redline, *underload*, leads to strong reductions in performance. In the current study, we focus on the levels of MWL around the right redline, i.e., the part of the curve where performance goes down. We coin this area the MWL upper red zone since these redlines rather represent a fuzzy zone when applied over contexts and people (cf., Young et al., 2014). We do not focus on performance decrements due to underload, i.e., lower red zone MWL (for an elaboration on underload and its relation to MWL levels, see Brookhuis & de Waard, 2000).

Within upper red zone MWL, a differentiation can be made between *impending* mental overload and mental overload per se. The MWL-redline approximates the break point (or a ceiling effect in terms of a limited resource system; Young et al., 2014) between both (see also, Grier et al., 2008; Paras, Yang, Tippey, & Ferris, 2015). We define impending mental overload as high MWL associated with minor decrements in performance and visualize this concept as our first Area Of Interest ('AOI 1') on the curve in Figure 1. Mental *overload* is defined as high MWL associated with major decrements in performance and is represented by the second Area Of Interest ('AOI 2') on the curve. As the curve shows an exponential trend, overload should thus be associated with the largest and most severe drops in performance.

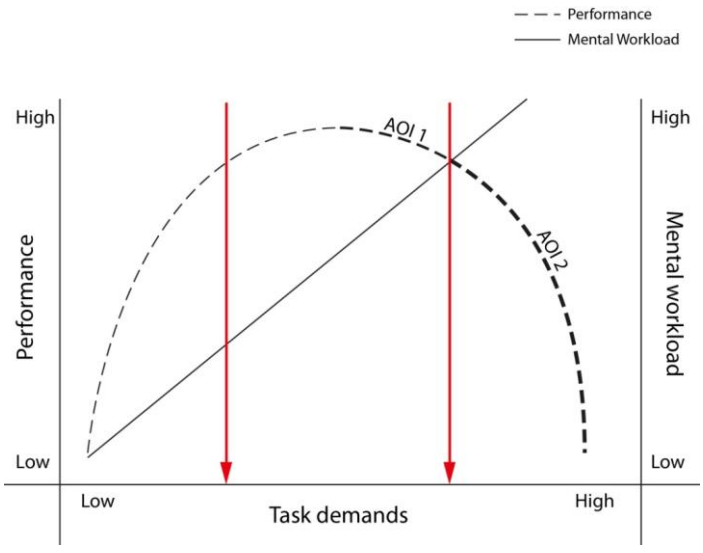

Figure 1: Graph of performance and MWL as a function of task demands, presenting the MWL-redlines and the areas of interest (AOI 1= Area of Interest 1, AOI 2= Area of Interest 2) . Adapted from Young and colleagues (2014) and de Waard (1996).

In all, we thus focus on red zone MWL associated with minor and major performance decrements – i.e., stretching over AOI 1 and AOI 2 in Figure 1. Our assumption subsequently states that performance decrements here will be observable in concrete assembly behavior and can hence indicate changes in MWL, in turn revealing changes in the cognitive resources spent. We illustrate this assumption in Figure 2. To facilitate theory building and comparability of the presented work (Van Acker et al., 2018), our operational definition of MWL states that *the MWL measured here will be reflected in the observed participants' assembly behavior indirectly revealing the interplay between participants' limited working memory modalities and the visual–spatial demands and working memory span demands being exposed to.*

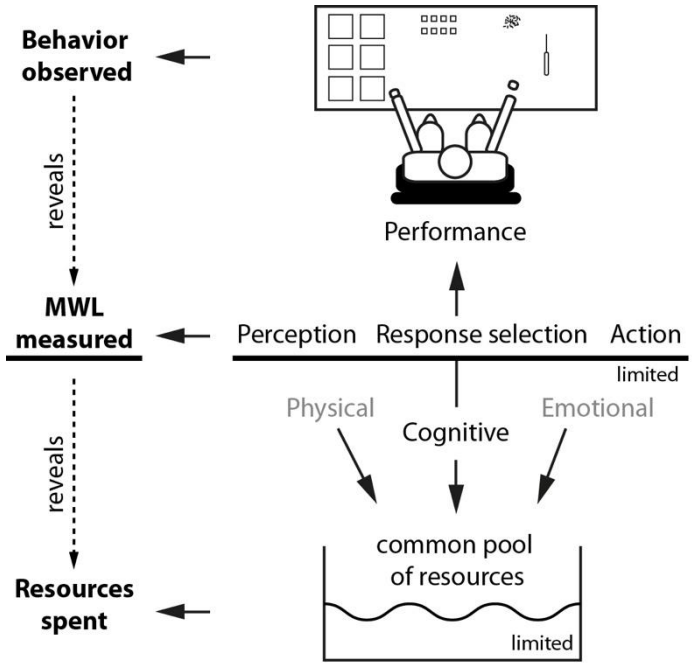

Figure 2 : Schematic overview of rationale how observable behaviors indicate MWL levels and thereby reveal the latent, non-observable competition for, and expenditure of, physiological resources.

3. Study overview

After delineating our focus area, we formulated our central research question stating: 'Which could be possible observable behavioral indications of upper red zone MWL?'. Based on Heyman and colleagues' (2018) recommendations, we tackled this question in twofold. We first developed a coding scheme identifying such behavioral indications

based on literature (deductively) and by observing assembly behavior in a subsample of videos from an assembly experiment validated to induce high and low MWL (inductively). After finalizing the coding scheme, we, secondly, validated the scheme with a larger set of videos of the same experiment (excluding the subsample used in the first stage). Specifically, we analyzed whether all behavioral codes would significantly differ per experimental condition (i.e., high and low complexity inducing high and low MWL) in their occurrence and, following Richardson and colleagues (2006), their duration.

3.1 Experimental procedures

For the experiment providing the video data for developing and validating the coding scheme, the same procedures as an earlier experiment exploring mobile pupillometry was followed (see Van Acker et al., 2020, for an extensive overview). In a within-subjects design experiment, participants performed both a high and low complexity assembly in a random and counterbalanced order and with a resting phase in-between. They assembled while being seated in a quiet room and wearing the SMI Eye-Tracking Glasses (ETG 2w) (SMI; Teltow, Germany), a lightweight glasses-type eye-tracking system (47g; size: 173 x 58 x 56mm) using small cameras implemented in the frame of the glasses (and compatible with contact lenses). Data were stored on a customized smartphone connected to the ETG with a cable. The smartphone was kept in the participants trouser pocket or waist bag. Participants were told that they would participate in a study 'exploring how people assemble and that their eye-movements would be tracked. They were explicitly instructed that they had all the time they needed to complete the assembly. In total the experiment would last for approximately one hour.

Both ecologically valid assemblies consisted of seven steps each, to be executed in a fixed order. Participants saw instructions per step on a screen at the opposite side of the worktable. Our focus was not on this instruction phase. After seeing the instructions, participants turned around to select, position and mount two components per step (three for the first step) that they had to select out of a display of parts presented on the worktable. After completion of each assembly, participants filled out a questionnaire. Figure 3 shows a diagrammatic overview of the assembly process.

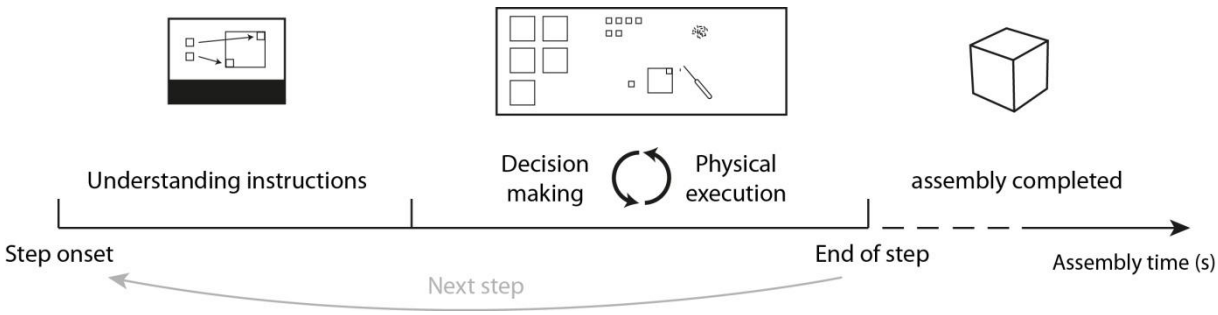

Figure 3: Diagrammatic overview of the assembly process. The low complexity assembly is used as an example. The high complexity assembly followed the exact same procedure.

The assemblies were designed to induce low and high MWL based on 10 task variables defining assembly complexity as proposed and validated by Richardson et al. (2004, 2006), Pillay's (1997) use of non-familiar assembly objects (see also Norman's, 1983, Shalin et al.'s, 1996, insights on mental models in objects) and one material-related task variable introduced by the authors. Specifically, the following 10 task variables defined the, respectively, low and high complexity level of the assemblies: assembly model (cubicle box design vs. abstract unfamiliar design, respectively), amount of symmetrical planes (high vs. low, respectively), amount of components (low vs. high, respectively), component groups (low vs. high, respectively), selections of components (no redundant components and displayed in correct order vs. redundant components and displayed in randomized order, respectively), amount of fastenings (minimally required vs. high, respectively), amount of fastening points (low vs. high, respectively), amount of novel (sub-)assemblies (low vs. high, respectively), presenting orientation on work table (correct angle vs. randomly angled 90° to the left or right along the horizontal plane, respectively), material of sides (low transparency vs. high transparency, respectively). The original experiment (see Van Acker et al., 2020) showed that subjective MWL differed significantly between the low and high complexity condition, in that the high complexity condition induced higher MWL levels as compared to the low complexity assembly (see Van Acker et al., 2020). The same assemblies used in the current experiment were thus validated to induce low and high MWL.

3.2 Video capturing procedures

The frontal Scene Camera of the eye-tracking glasses (with a resolution of 1280x960p; 24 FPS; H.264 video format, 60° horizontal and 46° vertical field of view; High Dynamic Range mode with high sensitivity for low light) was used to capture participants' assembly

behavior, i.e., their hands and upper arms manipulating the assembly. Videos (in *.avi* format) were then imported into the SMI BeGaze software (SMI; Teltow, Germany) in order to calculate the eye-gaze position (3-point calibration; binocular eye tracking at 60Hz with automatic parallax compensation and dark pupil tracking; 80° horizontal and 60° vertical gaze tracking range; gaze position accuracy of 0.5° over all distances) to be used for the definitions of certain behavioral codes.

4. Development of the coding scheme

4.1 Method

Behavioral observation is a method in which a researcher sees and/or hears, and then systematically records specific behaviors of an individual or group within a certain context of interest (Heyman, Lorber, Eddy, & West, 2018). These behaviors are then organized into categories by using clearly defined codes to be assigned based on certain rules. Together, these categorized codes form a coding scheme. After analyzing these behaviors from, for example, video data, their occurrence or time duration can be analyzed. No behavioral coding scheme for MWL measurement in assembly execution existed already. We - three of the current authors and a team of three trained graduate students - therefore developed a scheme through *deduction* and *induction*, inspired by the practical guide for developing behavioral coding schemes of Chorney and colleagues (2015).

4.1.1 Deductive path

From theory, we started with a profound understanding of MWL and how assembly behavior is propelled by mental processing (cf., the Background section). Most central here is that cognitive processing competes with physical processing for the limited physiological resources. The MWL measurement literature learns that assembly-specific cognitive processing can subsequently be inferred from execution time (Richardson et al., 2006; Stork & Schubö, 2010), reaction time and errors (Stork & Schubö, 2010; Young et al., 2014), dwell time on assembly and amount of sub-steps needed (Stork & Schubö, 2010), and even hand movement parameters such as total movement time, speed-accuracy tradeoff, peak velocity and latency of movement onset (Stork & Schubö, 2010).

Other work outside assembly environments infers MWL levels based on behavioral parameters such as writing velocity and pen gesture (Badarna, Shimshoni, Luria, & Rosenblum, 2018; see also Meulenbroek, Van Galen, Hulstijn, Hulstijn, & Bloemsaat, 2005; for an overview, see Chen et al., 2016). Both strongly resonating with our work here, Qiu and Helbig (2012) used video-analysis to show how computer operators' task-unrelated body posture (coming closer or moving less) can indicate MWL, whereas Boer (2000) proposed a behavioral entropy index characterizing MWL in driving (i.e., a quantification of corrective actions such as error corrections or startled responses reflecting the level of smoothness in driving control behavior such as lane keeping and car following). Next, features of computer mouse behavior such as contemplation-style pauses occur more and longer under high MWL (Arshad, Wang, & Chen, 2013; see also target highlight time in Vitense, Jacko, & Emery, 2003) – resembling the use of more and longer pauses when speaking under high MWL (Ruiz, Taib, & Chen, 2006) or, oppositely but revealing the interplay of resources, resembling poorer performance on a difficult secondary cognitive task while walking (Srygley, Mirelman, Herman, Giladi, & Hausdorff, 2009).

Based on these notions and in relation to the assembly stages for commissioning, joining and action execution, we first derived codes defined by execution time (e.g., the time to find the right position) and erroneous behavior (e.g., the amount of wrong positioning attempts). Based on, i.a., behavioral entropy, body posture and pauses, we distilled generic behavioral codes, such as muscular freezing (reflecting thinking or startle responses) and changes in the relative position of the head. A first preliminary coding scheme was thereby crafted, importantly, before seeing any videos.

4.1.2 Inductive path

In a next phase, we used the first coding scheme to code three training samples, being three videos selected from the collected videos of the same experiment, but not used for the validation. By using the behavioral video analysis software BORIS (Friard & Gamba, 2016), we now inductively refined, selected and pragmatized our first set of codes, and aimed at discovering new codes. For the latter, team members were instructed to identify additional exemplary behavioral indicators of MWL being too high, i.e., not yet captured in the first coding scheme. Specifically, they were instructed to gauge how

decrements in performance (quantified by duration and number of occurrences) manifested in observable behaviors other than those already identified in the deductive phase.

For this inductive method, team members alternated between working individually and discussing all adaptations in group. We discussed (also with team members not having been involved in the early stage iterations), tested and updated several coding schemes during multiple iteration rounds on completeness (i.e., do we agree that all possible behavioral indicators are included), accuracy and face validity (i.e., are the definitions adequate to grasp and delineate the behaviors aimed for), and feasibility (i.e., is the scheme usable for the raters, e.g., in terms of comprehensibility and vigilance).

4.1.3 Inter-rater reliability

The final coding scheme was fine-tuned during the last stages of the iteration process. Procedures for this were largely based on Lombard, Snyder-Duch and Bracken (2002, 2010) and McAlister and colleagues (2017), and went as follows. First, three team members used BORIS to independently code parts of the three training videos (not used for the validation) and repeatedly assessed the inter-rater reliability (IRR) informally by reflecting on their coding disagreements in group. Specifically, they discussed where and why in the videos they seemed to disagree. Subsequently, they redefined the respective codes until adequate agreement on the coding scheme was reached.

In a next stage, as is common practice in behavioral coding research (see also Hallgren, 2012; Neuendorf, 2002), two raters (i.e., different team members) consolidated the coding scheme and now formally assessed IRR based on a subset of four videos of the full sample to be used for the validation (for a similar approach, see, e.g., Baranek et al., 2005). Specifically, they first coded the videos of the first participant of the subset independently from each other and afterwards calculated the Cohen's Kappa IRR coefficient in BORIS (a robust standardized statistic for agreement between raters ranging from -1 to +1, with 0 being agreement expected by from random chance, and correcting for random agreement due to guessing, Cohen, 1960) with an interval time of 1.00s (meaning that the behaviors coded by the two raters were checked for

agreement/disagreement per second). Then, they discussed the minor unclarities remaining in the coding scheme as revealed by the IRR-calculation - for instance due to an atypical assembly strategy exhibited by the participant. When a consensus was reached on refining these subtle ambiguities accordingly and when the codes were thoroughly understood by the raters, the same videos were coded again from the beginning, independently and without guidance. IRR was then calculated again. This process of training and consolidation continued for the first and, if necessary, for the following participant videos until an acceptable Cohen's Kappa IRR coefficient was reached of $K > .70$ for the complete subset (falling within the range for 'moderate agreement' of 0.60 to 0.79, McHugh, 2012; we opted for this threshold being appropriate for a first exploratory study on the topic, see Lombard et al., 2002, 2010).

4.1.4 Coding procedures

The following coding approach was used throughout the process of informal (i.e., the three training video's) and formal IRR assessment (i.e., the subset of four), and throughout the validation process (see below). Out of the, in total, seven steps both assemblies consisted of, only the first, third and fifth step of both assemblies were coded in order to make the labor-intensive video coding process feasible for the raters while selecting a representative sample of assembly behavior (for a similar approach, see, e.g., Baranek et al., 2005). This provided in total approximately 10 minutes of data per participant[1] to be analyzed with BORIS. Specifically, a combination of a topographical coding system measuring the occurrence of behaviors (i.e., the amount of times a code occurred) and a dimensional coding scheme gauging the time duration of behaviors was deployed (Heyman et al., 2018). For the detection of occurrence and duration, a timed-event sequential continuous coding approach was used (Chorney et al., 2015). That is, the raters continuously analyzed behavior during the complete assembly step by backwarding and forwarding throughout the videos.

Although preferred when possible, raters could not be blind for the manipulation, since the design of the assemblies were also visibly different in complexity for the raters. Since

[1] Note that participants had all the time they needed to complete each step, so that the length of data differed per participant.

the codes are clearly delineated in time and define concrete physical behavior (instead of, e.g., socially constructed behaviors, see Chorney et al., 2015), we do not foresee the coding having been prone to judgment subjectivity. Table 1 provides an overview of the steps and amount of participants coded per rater.

	Assembly steps							
	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	
Rater 1	x		x		x			*N* = 24
Rater 2	x		x		x			*n* = 4

Table 1: Overview of assembly steps coded for both assemblies and the amount of subjects coded per rater. "X" indicates that a step was coded by a rater.

The final coding scheme differentiated codes between state events (for which both duration and occurrence was important) and point events (for which duration was not meaningful and only occurrence was hence coded). For the state events, onset and stop times were coded. Point events were coded at the onset of the behavior. For the onsets and stop times, both raters aimed to apply an accuracy of approximately 0.5 seconds. Not every behavior was coded. Specifically, we did not account for ergonomic behavior (i.e., repositioning one's hand, retaking a part for comfort or picking up a part the participant dropped), orderliness behaviors such as putting parts aside in a personally preferred way, screwing or picking screws. We did so, since we did not have deductive nor inductive grounds for linking such behaviors to upper red zone MWL. Figure 3 shows a screenshot of the behavioral video coding software deployed to apply (iterations of) the coding scheme.

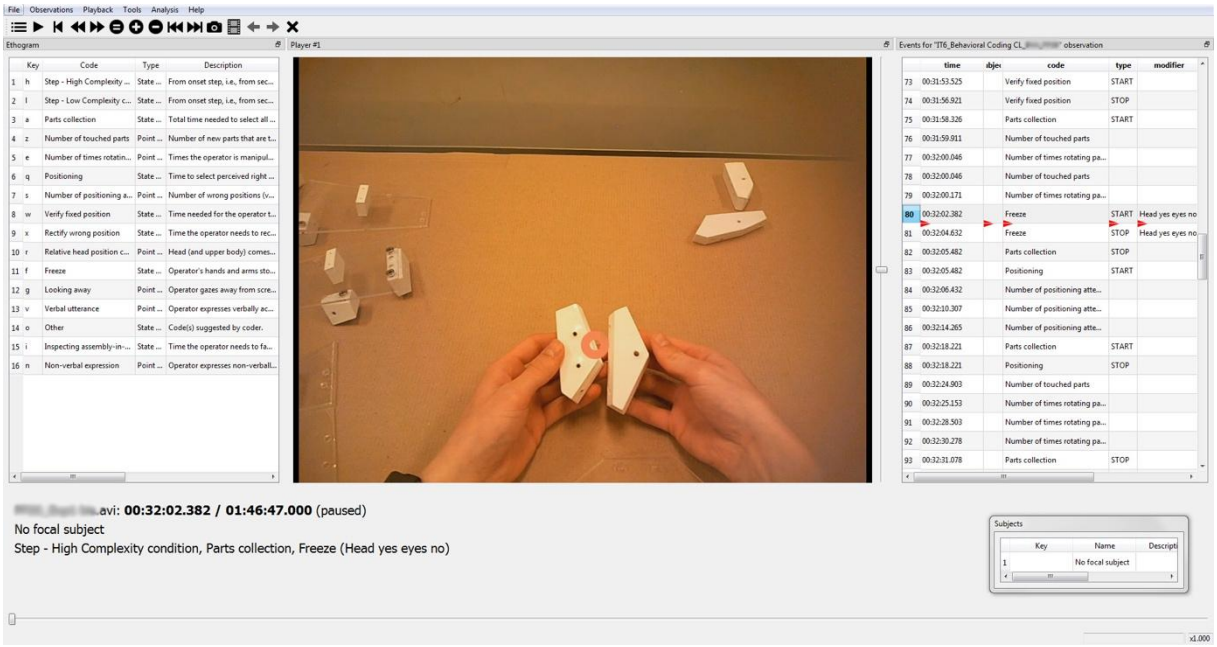

Figure 4: Screenshot of the coding scheme (left side of the screen, see 'Ethogram') deployed in BORIS, on a participant video (center) (orange circle displays the eye gaze), resulting in a sequence of behaviors coded throughout the video (right side of the screen, see 'Events for …').

4.2 Results

The final coding scheme, consisting of 4 categories and 11 codes, is displayed in Table 1 and was synthesized in a logic sequence of categories for assembly execution, being; I., Parts selection behavior (3 codes), II., Positioning behavior (4 codes), III., Transition behavior (i.e., in-between completing an assembly step and starting the next) (1 code) and, IV., Generic behavior (occurring throughout all assembly phases) (3 codes).

Only the code 'Freeze' involved modifiers to differentiate between three types of freezes. Importantly, we used the participant's eye-gaze point to distinguish between certain codes and to help define others. Table 1 also includes the available empirical referents that helped in the coding scheme development. Note that some codes were mainly driven by the inductive path, others deductively. In Appendices we included a table with rater instructions (Appendix A) and one with rater attention points (Appendix B), both especially to be used during rater training.

Finally, not all codes were mutually exclusive, meaning that some codes could be nested in another code, such as 'Gaze redirection' in 'Freeze', or 'Freeze' in 'Verification of fixed

position'. Other codes were always nested in the main categorical code, such as 'Part touched' in 'Parts collection'.

***Category*, Code (Unit)**	**Definition**	**Empirical referents**	**Example**
I. Selecting Behavior			
1. Parts collection (state event)	Time needed to select all perceived correct parts.	Execution time (Richardson et al., 2006; Stork & Schubö, 2010) and reaction time (Young et al., 2014)	The operator wants to select the required two parts, but under higher MWL levels touches five parts and rotates them several times before a decision is made on which the correct parts are. In total, the selection of both parts hence lasted longer.
2. Part touched (point event)	Number of new parts touched.	Making errors (Richardson et al., 2006; Stork & Schubö, 2010), amount of sub-steps needed (Stork & Schubö, 2010)	
3. Part rotation (point event)	Number of times operator is manipulating a part (here rotating, from velocity is 0 to velocity is 0).	Making errors (Richardson et al., 2006; Stork & Schubö, 2010), amount of sub-steps needed (Stork & Schubö, 2010)	
II. Positioning Behavior			
4. Positioning (state event)	Time to select perceived right position. Includes rotation and alignment.	Execution time (Richardson et al., 2006; Stork & Schubö, 2010) and reaction time (Young et al., 2014)	When then positioning the two parts, the operator, still under higher MWL levels, makes a few redundant positioning attempts, needs more time to verify whether a fixed position is then indeed correct and takes a while to correct a wrong perceived fixed position. Altogether, positioning the two parts lasted longer as would be expected.
5. Positioning attempt (point event)	Number of positions tried (defined by velocity = 0, part touches assembly-in-progress or is being held above, below or aside from it) on assembly-in-progress.	Making errors (Richardson et al., 2006; Stork & Schubö, 2010), amount of sub-steps needed (Stork & Schubö, 2010)	
6. Verification of fixed position (state event)	Time needed for operator to verify if a position (i.e., orientation, alignment, position) is correct after a component has been fastened, for at least 2 seconds.	Dwell time on assembly (Stork & Schubö, 2010)	
7. Correction of perceived wrong position (state event)	Time operator needs to correct a (set of) perceived wrong fastening(s) and perform a perceived correction, for at least 2 seconds.	Error corrections in behavioral entropy (Boer, 2000)	

III. Transition Behavior			
8. Inspection of assembly-in-progress (state event)	Time operator needs to familiarize with or verify the assembly-in-progress including touching and rotating assembly-in-progress, for at least 2 seconds, in between visual intake of instructions and onset 'Parts collection' or 'Correction of perceived wrong position'.	Dwell time on assembly (Stork & Schubö, 2010)	When proceeding to the next step, the operator first inspects the assembly-in-progress. Under higher MWL levels, this lasts longer.
IV. Generic Behavior			
9. Freeze (state event)	Number of times both operator's hands and arms stop moving or only move very minimally (the latter not as a function of selection or positioning), at a fixed place in the workspace for at least 2 seconds.	Hand movement parameters (Stork & Schubö, 2010), writing velocity (Chen et al., 2016), startled response in behavioral entropy (Boer, 2000), contemplation-style pauses in computer mouse behavior (Arshad et al., 2013) and dwell time on assembly (Stork & Schubö, 2010)	During the entire assembly process, apart from the assembly stages, the operator also freezes more often, has several occasions of needing to redirect the eye gaze and more often inspects the assembly (parts) from closer by – all when under higher MWL levels.
Modifier 1	Head freezes too, eyes move		
Modifier 2	Not accompanied by freeze of head and eyes		
Modifier 3	Both eyes and head freeze too		
10. Gaze redirection (point event)	Number of times operator gazes away from worktable for at least 2s.	Behavioral entropy (Boer, 2000)	
11. Relative head position change (point event)	Number of times operator's head is repositioned closer to or further away from parts or assembly-in-progress, because of the operator moving the upper body or oppositely, the parts or assembly-in-progress.	Task-unrelated body posture (coming closer, moving less) (Qiu & Helbig, 2012)	

Table 2: Final behavioral video coding scheme.

5. Validation of the coding scheme

To validate our central research question on possible observable behavioral indications of upper red zone MWL, we deployed our coding scheme on a set of 24 videos from the assembly experiment and hypothesized that:

> *H: The behaviors detected will last longer and/or occur more frequently in the high complexity condition as compared to the low complexity condition.*

We directly derived this hypothesis from previous work empirically addressing the limited resource theories (discussed in section 2. Background) by showing that, e.g., higher behavioral entropy in driving and more and longer contemplation-style pauses in computer mouse behavior indicate high MWL (see section 4.1.1). The experiment leveraging a high and a low complexity assembly now served to validate, with a separate large-scaled sample of videos, whether the behavioral codes of the coding scheme are indicative of high MWL, as compared to low MWL. This way, we experimentally tested the coding scheme's discriminative validity, i.e., can the behavioral codes distinguish between groups that are hypothesized to differ (Heyman et al., 2018)?

5.1 Participants

Because no previous studies exist deploying behavioral video coding to gauge MWL, we looked for previous work using the most similar research methodology to help us determine the sample size. The work of Qiu & Helbig (2012) and Baranek and colleagues (2005) adhered most closely to our experimental setup. The former inferred MWL levels through automated video analysis of body posture based on 22 participants and a pair-wise comparison between cognitive tasks in a within-subjects experimental design. The latter study by Baranek and colleagues (2005) gauged autism in infants by deploying a retrospective video analysis method very similar to our method. In this study, 32 participants were distributed over three conditions in a between-subjects experimental design. In line with these most similar studies, and to keep the labor intensive video coding feasible, we aimed to also arrive at a minimum of 20 participants per condition in our within-subjects design experiment. A total of 25 university student volunteers of an engineering university faculty (20% female, M_{age} = 21.48, SD_{age} = 1.19), naïve to the manipulation participated after giving a written consent (no participant afterwards reported to have an idea about the experimental goal). Participants' inclusion criteria were the same as in Van Acker and colleagues (2020).

5.2 Measures

As the experiment now serves for validation of the codes, we first checked whether the manipulation of the experiment was successful in inducing MWL and excluding confounders by including subjective measures and a test. The same measures and their

theoretical background were used as in earlier work (see Van Acker et al., 2020, for more information). All items deployed a 7-point Likert-scale.

5.2.1 Manipulation check

As a manipulation check, perceived task *Complexity* was measured with a single item ('I perceived this assembly to be difficult.') and subjective *Mental Workload* (MWL) was gauged by the average of the scores on three items derived from cognitive load theory (Paas, 1992; Paas, van Merriënboer, & Adam, 1994), the Subjective Workload Assessment Technique (SWAT; Reid and Nygren 1988) and the NASA-TLX (S. G. Hart & Staveland, 1988) (e.g., 'I experienced this assembly as cognitively demanding'; $\alpha_{\text{low complexity}} = .90$, $\alpha_{\text{high complexity}} = .91$).

Since we aimed to induce (impending) mental overload, we also measured emotional load (EL), as it can provide an indirect measure (Van Acker et al., 2018). *Emotional Load* consisted of five items selected from the Dundee Stress State Questionnaire (DSSQ; Matthews et al. 2013; Matthews 2016) (e.g., 'I felt frustrated while performing this assembly.'; $\alpha_{\text{low complexity}} = .77$, $\alpha_{\text{high complexity}} = .83$) measuring the negative emotions frustration and irritation, whether participants felt tense during the assembly, and appraisals covering whether they felt they could cope with the situation and to what extent they felt uncertain.

5.2.2 Measures of confounding variables

As we only aimed to measure (impending) overload, we wanted to minimize effects of other possible confounding variables. Therefore, we first measured perceived *Physical Load* (PHL) with one item ('I perceived this assembly to be physically demanding') as it can interact with mental and emotional load, hence confounding the duration of our behavioral codes (i.e., higher PHL inducing longer code durations because of physical fatigue). Additionally, two variables *Mind Wandering* and *Fed Up* consisting of one item each (also derived from the DSSQ, Matthews et al. 2013; Matthews 2016) checked for task engagement by asking whether the participant's mind started wandering and to what extent the participant was fed up with the assembly. We included these questions to check whether the coded behavior could be caused by a lack of engagement into the

assembly task (e.g., freezing longer because of mind wandering or more parts touched because of being fed up with the task).

Finally, two idiosyncratic control variables were included for which we reasoned they could correlate with our codes. First, we measured the self-reported *Dexterity* with one item ('In general, how dexterous do you estimate yourself to be - apart from how you experienced these assembly tasks'), again on a 7-point Likert answer scale (from 1, very clumsy, to 7, very dexterous). Second, participant's *visual-spatial intelligence* was gauged with a subset of the Revised Minnesota Paper Form Board test (Stinissen, 1977) subsequent to both experimental conditions and a resting phase of five minutes. We did so, since this intelligence factor can largely affect interpersonal differences in assembly performance.

5.2.3 Video coding procedures

As proposed by Hallgren (2012; see also Neuendorf, 2002), after calculation of the inter-rater reliability (see above), the primary rater continued to analyze the remainder of the dataset in BORIS using the final coding scheme and following the same coding procedures (for a similar approach, see Baranek et al., 2005). The secondary rater thus only served for reliability. In total, 24 videos were coded by this rater[2], i.e., the four videos already coded for the IRR assessment and 20 remaining videos. These 24 coded videos were used for the analysis. Note that the first, third and fifth step were selected to code. Below, we refer to these steps as Step 1, Step 2 and Step 3, respectively.

From BORIS, we subsequently extracted the occurrence and duration of the codes per step. As duration and occurrence was sometimes important per code, while for other codes only occurrence was relevant, we hence did not have 22 dependent variables in total, but 19. All codes measured through duration were quantified in seconds. All codes measured through occurrence were quantified as the number of times the coder had coded the behavior (e.g., 'Parts collection' sometimes occurred multiple times within a

[2] Note that we had 25 participants in total, but for one participant the video was corrupted and therefore not usable, because of a technical malfunction.

step, because the participant returned to a 'Parts collection' phase after starting 'Positioning').

5.3 Results

All subjective measures lacked normality of data. We therefore ran non-parametric tests. The results are reported in Figure 4.

5.3.1 Manipulation check

A Wilcoxon Signed Rank Test on perceived assembly difficulty showed that the high complexity condition was indeed perceived as more difficult ($M_{\text{high complexity}}$ = 4.84, *SD* = 1.56, range: 1-7; $M_{\text{low complexity}}$ = 1.56, *SD* = .92, range: 1-5; *N* = 25), *z* = -3.99., *p* < .001 (two-tailed), with a large effect size of *r* = .56 (i.e., $r = z/\sqrt{(Nx + Ny)}$, Rosenthal 1994). A Wilcoxon Signed Rank Test showed that MWL was perceived as higher in the high complexity condition with a rather neutral score ($M_{\text{high complexity}}$ = 4.39, *SD* = 1.58, range: 1-7; $M_{\text{low complexity}}$ = 2.10, *SD* = 1.12, range: 1.00-5.33; *N* = 25), *z* = -3.79, *p* < .001 (two-tailed) and with a large effect size of *r* = .54. The complexity manipulation was hence effective.

For EL, a Wilcoxon Signed Rank Test revealed that EL was significantly higher in the high complexity condition ($M_{\text{high complexity}}$ = 3.72, *SD* = 1.21, range: 1.60-6.00; $M_{\text{low complexity}}$ = 2.16, *SD* = 1.04, range: 1.00-4.80; *N* = 25), *z* = -3.71, *p* < .001 (two-tailed), with a large effect size of *r* = .53.

5.3.2 Confounding variables

Physical Load (PHL) showed to be higher in the high complexity condition ($M_{\text{high complexity}}$ = 1.72, *SD* = 1.06, range: 1-5; $M_{\text{low complexity}}$ = 1.16, *SD* = .55, range: 1-3; *N* = 25), *z* = -2.38, *p* = .018 (two-tailed), with a medium effect size of *r* = .34, but was still low.

Being Fed Up with the assembly was higher for the low complexity condition, expressed in a neutral score as compared to a low score for the high complexity condition ($M_{\text{high complexity}}$ = 2.40, *SD* = 1.32, range: 1-5; $M_{\text{low complexity}}$ = 4.00, *SD* = 1.89, range: 1-6; *N* = 25), *z* = -2.73, *p* = .006 (two-tailed), with a medium effect size of *r* = .39. Mind Wandering was rated higher for the low complexity condition as well, expressed in a neutral score, as compared to a low score for the high complexity condition ($M_{\text{high complexity}}$ = 2.44, *SD* = 1.53,

range: 1-6, N = 25; $M_{low\ complexity}$ = 4.08, *SD* = 1.91, range: 1-7, *N* = 24), *z* = -2.61, *p* = .009 (two-tailed), with a medium effect size of *r* = .37.

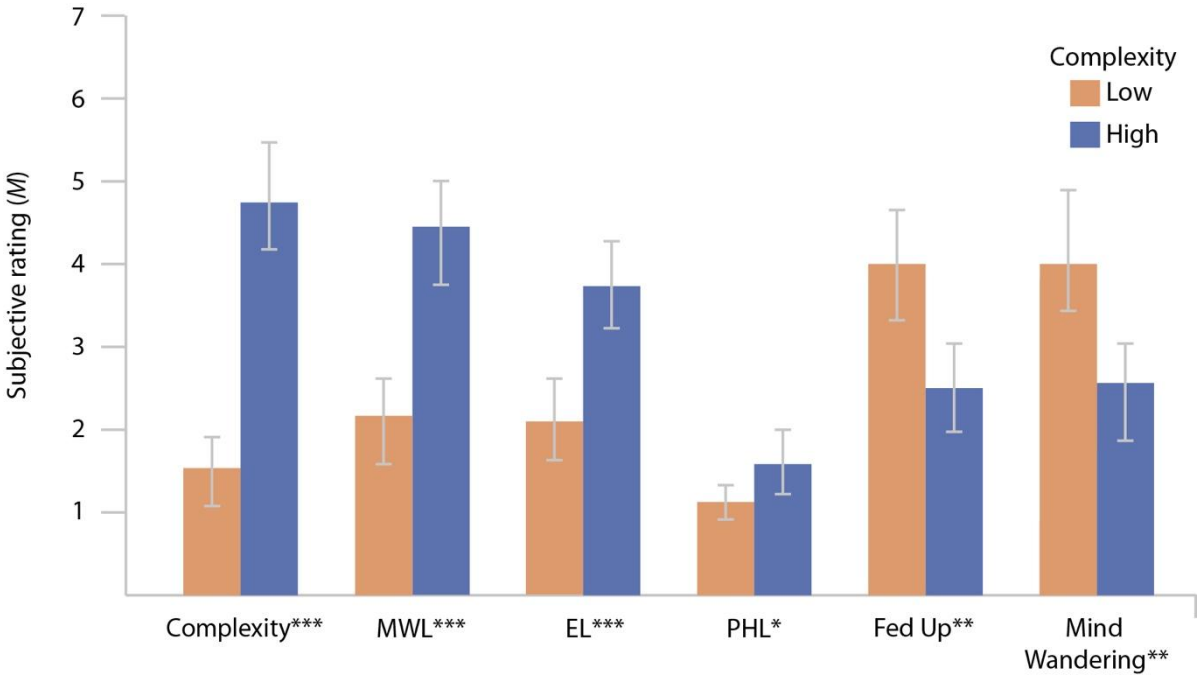

Figure 5: Overview of the means (with 95% CI) of all subjective ratings used for the manipulation check and possible confounding variables, **p* < .05, ***p* < .01, ****p* < .001. Note: MWL = mental workload; EL = emotional load; PHL = physical load; 'Complexity' refers to the experimental conditions.

Summarized, our manipulation succeeded in inducing two different levels of MWL, though during the high MWL condition levels still remained rather neutral. EL also differed, but the high complexity condition neither yielded high EL, hence not potentially indicating mental overload. We hoped being fed up and mind wandering to be low, but they panned out closer to average. These negative indications of task engagement differed per condition too, potentially confounding our behavioral coding data towards not finding differences between conditions (i.e., possible negative confounding because of slower and ineffective assembly behavior in the low complexity condition). Finally, physical load unexpectedly differed per condition, although it remained low to non-existent for both conditions (i.e., very close to the lower extreme of the scale). For the latter reason, we do not expect physical load to confound the occurrence and duration of our codes.

5.3.3 Behavioral coding

Contrary to our subjective measures where the unit of analysis was our participants, in our behavioral coding analysis, we used a flattened, nested dataset consisting of n = 144 coded observations across 24 participants[3], that is, a dataset of all occurrences and durations per code (i.e., the 19 columns for the dependent variables) for; the coded steps (i.e., 3), per condition (i.e., 2), per participant (i.e., 24) - making 144 rows of observations (i.e., 3 x 2 x 24 = 144).

As discussed by Snijders and Bosker (1999) approaches such as analysis of variance or general linear modeling assumes that no relationship exists between the individual observations in the sample. In our study, this is clearly not the case because multiple observations come from the same participant. As a result of this, there can be an assumption that the individual will impact our various outcome variables. We therefore used multilevel modeling (also called mixed effects or random-effects models) to test for the effects of the different codes, with the individual participant being the grouping variable. As discussed by Maas and Hox (2005), our sample size is sufficiently large to perform a multilevel model. We used the lme4 (1.0-5) R Package (Bates et al., 2019).

We first report a null model (I) including no variables, to show differences with the subsequent models, being a model (II) with only the control variables (i.e., spatial intelligence and dexterity), a model (III) with only the independent variables, a model (IV) with the independent variables and their interactions and a final full model (V) including all these variables. Per model, we report the total explained variance with an R^2 estimate. Because of the large amount of dependent variables, we included all respective tables in Appendices. Note that Step 1, Step 2 and Step 3 referred to below are steps selected from the seven assembly steps and therefore represent, respectively, steps 1, 3 and 5 of both assemblies (see Table 1).

Selecting behavior

For the full model (V), the code 'Parts collection' was found to indeed occur more (p < .001) and longer (p < .001) during the high complexity Condition, as compared to the low

[3] Because we did not have the spatial intelligence data for one participant, we ran the analysis with 24 participants to assure complete comparability.

complexity Condition. This full model explained a total variance of 31% (of which 24% by the independent variables).

This behavior also occurred more ($p < .01$) and longer ($p < .001$) during the first step, as compared to the second step. Step 2 (as compared to Step 1) interacted with Condition for the number of occurrences ($p < .01$) and duration ($p < .05$), so that during Step 1 this code was even more prevalent and took even longer during the high complexity Condition. Additionally, there was an interaction between Step 3 (as compared to Step 1) and Condition ($p < .01$) for the duration of this code in the same direction (see Figure 6). This full model explained a total variance of 48% (of which 43% by the independent variables).

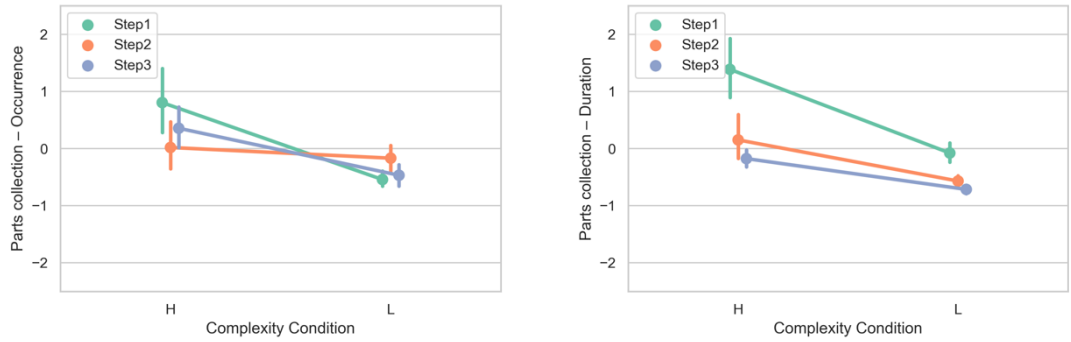

Figure 6: Occurrence and duration of 'Parts collection', expressed in z-scores.

'Part touched' did not differ per Condition. Again, the behavior occurred more during the Step 1, as compared to Step 2 ($p < .001$), and Step 3 ($p < .001$) (see Figure 7). Dexterity interestingly, was positively related to this code too ($p < .05$), although its explained variance was low ($R^2_{\text{model II}} = .03$, $R^2_{\text{model III}} = .24$, $R^2_{\text{model V}} = .25$).

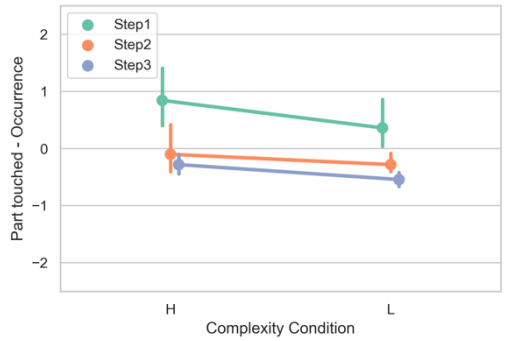

Figure 7: Occurrence of 'Part touched', expressed in z-scores.

'Part rotation' occurred more under high complexity levels, $p < .05$, and again more during Step 1, as compared to Step 2, $p < .001$, and Step 3, $p < .001$ ($R^2_{\text{model III}} = .31$, $R^2_{\text{model V}} = .31$) (see Figure 8).

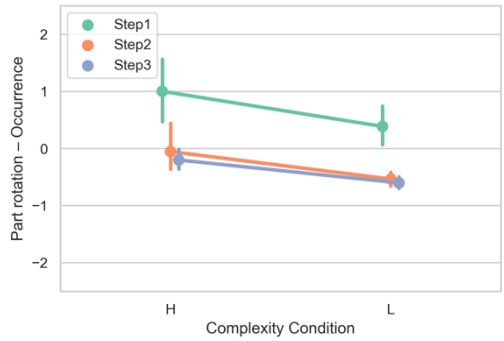

Figure 8: Occurrence of 'Part rotation', expressed in z-scores.

Positioning behavior

'Positioning' too, was more prevalent during the high complexity Condition, $p < .001$. An interaction effect between Condition and Step 2, $p < .05$, showed that the code occurred significantly more during Step 2 during the low complexity Condition as compared to Step 1 ($R^2_{\text{model III}} = .23$, $R^2_{\text{model V}} = .27$).

The duration of 'Positioning' also took longer during the high complexity Condition ($p < .001$) and, remarkably, during Step 3, as compared to Step 1 ($p < .001$). An interaction effect showed the behavior took longer during Step 3 (compared to Step 1) under high complexity levels, $p < .01$ ($R^2_{\text{model III}} = .48$, $R^2_{\text{model V}} = .54$) (see Figure 9).

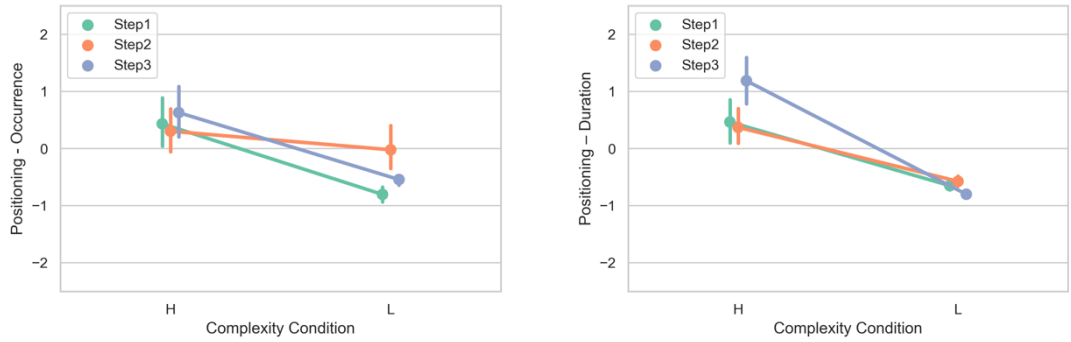

Figure 9: Occurrence and duration of 'Positioning', expressed in z-scores.

'Positioning Attempt' was observed more under high complexity levels, $p < .001$, and for Step 1 as compared to Step 2, $p < .05$ ($R^2_{\text{model III}} = .30$, $R^2_{\text{model V}} = .32$) (see Figure 10).

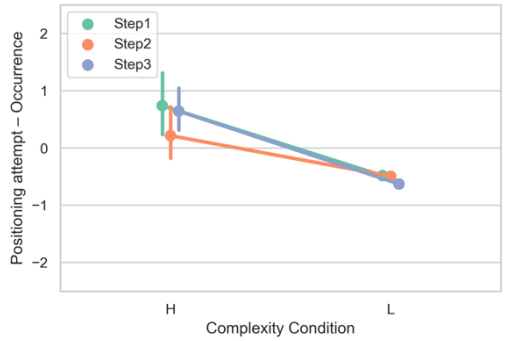

Figure 10: Occurrence of 'Positioning attempt', expressed in z-scores.

'Verification of fixed position' did not depend on Condition, but did occur more during the first step as compared to Step 3, $p < .01$ ($R^2_{\text{model III}} = .22$, $R^2_{\text{model V}} = .24$).

The duration of this code did not depend on Condition. Again, the code lasted longer during Step 1 when compared to Step 3, $p < .05$ ($R^2_{\text{model III}} = .27$, $R^2_{\text{model V}} = .30$) (see Figure 11).

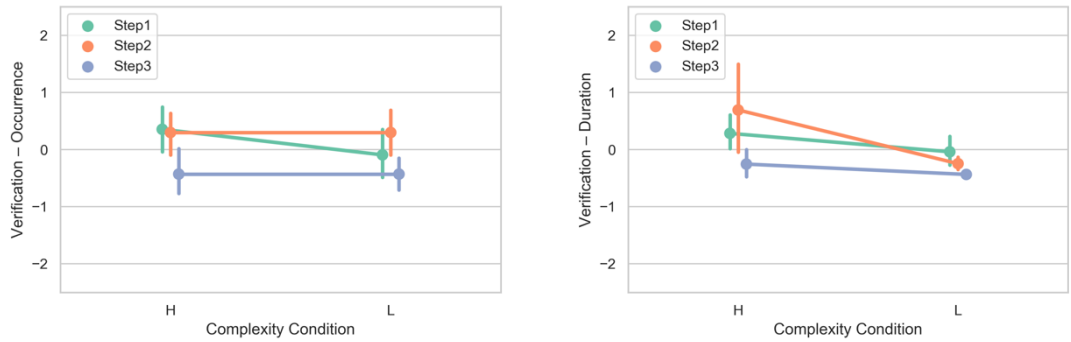

Figure 11: Occurrence and duration of 'Verification of a position', expressed in z-scores.

The amount of occurrences of 'Correction of perceived wrong position' was not observed differently per Condition, but it did occur more during Step 2 as compared to Step 1, $p < .001$, and showed an interaction effect, $p < .001$, showing that the code occurred more often during Step 2 under high complexity levels when compared to Step 1 ($R^2_{\text{model III}} = .18$, $R^2_{\text{model V}} = .27$).

Nor did the duration of this behavior show to depend on Condition. The duration was however longer during Step 2 ($p < .01$) and Step 3 ($p < .01$), both as compared to Step 1. An interaction effect additionally showed that the code lasted longer during Step 2 ($p <$

.05) when compared to Step 1, during the high complexity Condition (see Figure 12). Finally, also Dexterity positively affected this code, $p < .05$ ($R^2_{\text{model III}} = .21$, $R^2_{\text{model V}} = .24$).

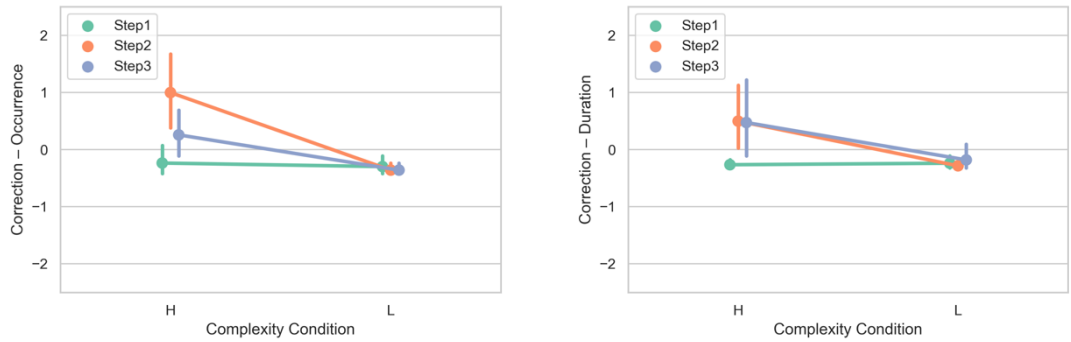

Figure 12: Occurrence and duration of 'Correction of a wrong position', expressed in z-scores.

Transition behavior

For 'Inspection of assembly-in-progress' we only compared Step 2 with Step 3, because during Step 1 no assembly-in-progress was yet present. This behavior occurred more during the onset of Step 3 as compared to Step 2, $p < .01$, but was independent of Condition. An interaction effect however revealed that this behavior occurred more during Step 3 under high complexity levels as compared to Step 2 (see Figure 13). Dexterity was also positively related to this behavior, $p < .05$, but only added little variance ($R^2_{\text{model III}} = .07$, $R^2_{\text{model IV}} = .17$, $R^2_{\text{model V}} = .20$). For the duration of this code, no effects were found ($R^2_{\text{model III}} = .6$, $R^2_{\text{model IV}} = .11$, $R^2_{\text{model V}} = .20$).

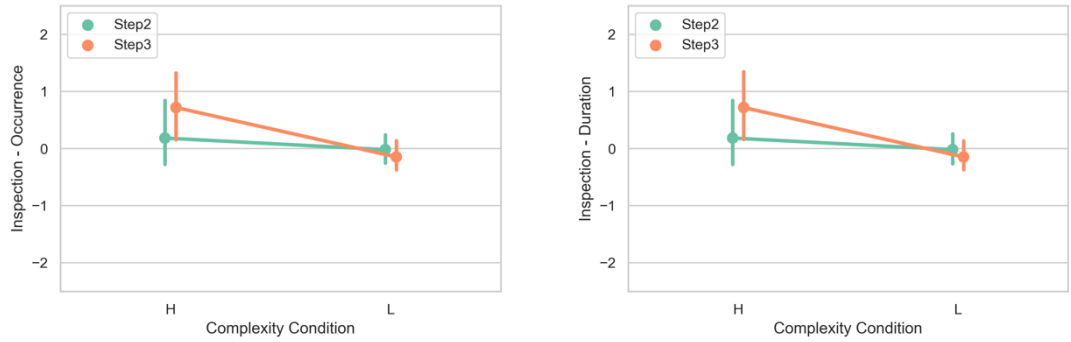

Figure 13: Occurrence and duration of 'Inspection of the assembly-in-progress', expressed in z-scores.

Generic behavior

'Freeze - Modifier 1' – did not depend on Condition ($R^2_{\text{model III}} = .16$, $R^2_{\text{model V}} = .19$) after adding the interactions in model III. Its duration was independent of Condition, as well

after adding the additional variables in model III. During Step 3 ($p < .05$) in general, this behavior however lasted longer, as compared to Step 1 ($R^2_{\text{model III}} = .16$, $R^2_{\text{model V}} = .20$). Additionally, an interaction effect, $p < .05$, revealed that this behavior lasted longer during Step 3 under high complexity levels, when compared to Step 1 (see Figure 14).

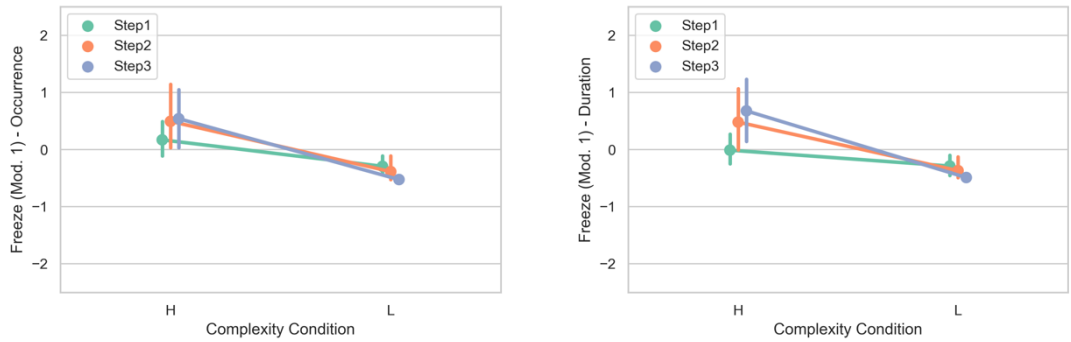

Figure 14: Occurrence and duration of 'Freeze (Modifier 1)', expressed in z-scores.

'Freeze - Modifier 2' occurred more during the high complexity Condition ($p < .05$) ($R^2_{\text{model III}} = .17$, $R^2_{\text{model V}} = .17$). The effect of Condition on its duration disappeared after adding the interactions in Model III ($R^2_{\text{model III}} = .10$, $R^2_{\text{model V}} = .10$) (see Figure 15). 'Freeze - Modifier 3' was only observed once, so that no analyses could be run.

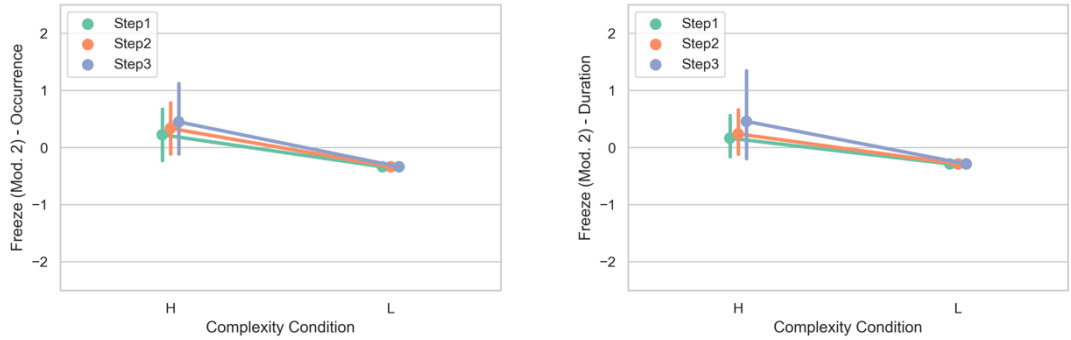

Figure 15: Occurrence and duration of 'Freeze (Modifier 2)', expressed in z-scores.

'Gaze redirection' did not depend on Condition, nor the Steps (see Figure 16). Only Dexterity showed a positive correlation ($p < .05$). The total explained variance of the full model (V) was nevertheless low ($R^2_{\text{model V}} = .06$).

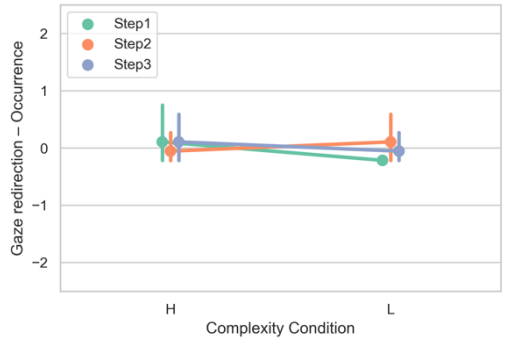

Figure 16: Occurrence of 'Gaze redirection', expressed in z-scores.

Finally, also 'Relative head position change' did not differ per Condition (see Figure 17). The total explained variance of the full model (V) for this code was also low, only explaining 5% the total variance (of which 4% by the IV's).

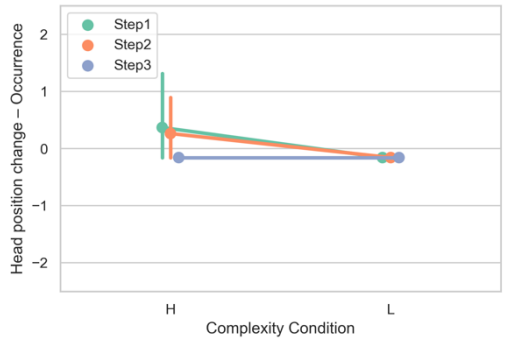

Figure 17: Occurrence of 'Relative head position change', expressed in z-scores.

6. Discussion

The current work focused on the 'Employee Work Behavior'-level as presented in the conceptual MWL-framework of Van Acker and colleagues (2018), thereby focusing on the underrepresented behavioral component of MWL. Iterative processes of deductive and inductive development were finalized into a behavioral video coding scheme defining 11 observable indicators of behaviors potentially revealing upper red zone MWL in assembly. An experiment was deployed to validate these codes, i.e., to answer the question whether these codes indeed measure what they should measure.

Multilevel analyses showed that only the occurrence and duration of the codes 'Parts collection' and 'Positioning', and the occurrence of 'Part rotation', 'Positioning attempt'

and 'Freeze (Modifier 2)' statistically significantly differed between conditions - thereby partly confirming our hypothesis. The behaviors not differing significantly, were; 'Part touched', 'Verification of a fixed position', 'Correction of a wrong position', 'Inspection of the assembly-in-progress', 'Freeze (Modifier 1)', 'Gaze redirection' and 'Relative head position change'. Based on this first presented work, five behavioral codes (out of eleven) can thus be withheld as indicative of upper red zone MWL in assembly.
In other words, these results suggest that upper red zone MWL can be observed as expressed during a complex assembly task in specific behavioral manifestations such as slower and/or more (1) collecting, (2) positioning, (3) rotating and (4) positioning attempts of/with assembly parts, and (5) freezing of the hands and arms.

The six non-significant codes that could not be indicative of MWL, in this context, could use more scrutiny or might reflect that behavioral expression of upper red zone MWL is partly idiosyncratic and might differ substantially between human beings. Indeed, some people might express different behaviors indicating fluctuations in MWL than others, so that for some certain behaviors might be very indicative, while other behaviors would not, or very rarely, occur at all. Also, it might be that upper red zone MWL elicits in some people rather covert human reactions (e.g. emotional, cognitive rumination) than the overt behaviors we aimed to observe. Below, we refer to some exploratory suggestions on personality traits in this respect. Next, the participant sample size might not have been sufficient to find all effects expected in this more real-life experiment. In line with this, it might also have been the case that the desired effects were not present as strongly as expected. As we show and elaborate upon below, most participants only experienced moderate levels of MWL. Under higher levels of MWL also some of the now non-significant behaviors might therefore become indicative of MWL as well. Inadequate task manipulation and raters' lack of familiarity with behavioral coding or judgmental bias as potential explanations for our non-significant findings can be excluded as we checked our experimental task manipulation and the raters involved were experienced senior researchers, well instructed and trained.

Interestingly, we did see that some main effects of Step might allude to some codes rather revealing accumulated cognitive fatigue towards the third step (Step 2 in the

analyses) and the fifth step (Step 3 in the analyses) out of the seven in total (for a possible example of such fatigue effects in pupillometry, see Van Acker et al., 2020; also see Hockey, 1997), whereas for other codes main effects could point to a more stressful, startled response-like effect at the onset of the assembly. The interaction effects of certain steps could nevertheless point in the same direction of our hypothesis, in that some codes occurred more and longer during the high complexity condition, but only for certain steps. Overall, main effects of Steps, the interaction effects found in the full models and the main effects of Condition disappearing after the interaction effects were included (in model III, cf., Freeze – Modifier 1) suggest that the complexity manipulation was not equally strong per step. The steps thus also varied in the MWL they induced. This is intelligible, since manipulating complexity in a real-world assembly with consecutive steps entails this challenge (also see Van Acker et al., 2020). Altogether, all codes might also have been underestimated because of negative confounding due to the low complexity condition yielding moderate levels of task engagement (as compared to significantly higher task engagement levels during the high complexity condition).

Of our control variables we found that spatial intelligence did not show any effect and dexterity only very weakly (see the modest amounts of explained variance). Exploratively, we were able to also measure three personality traits for which we thought they could relate to our codes, this for a subsample of $n = 17$. We subjectively gauged 'Impulsivity', i.e., the tendency of acting before thinking in work situations (based on 2 items measuring 'individual action propensity' in Vera, Crossan, Rerup, & Werner, 2014), and we gauged the Big Five personality traits *Neuroticism* and *Conscientiousness* (each measured with 4 items out of a short form of a Big Five scale; Donnellan, Oswald, Baird, & Lucas, 2006). We found that only Impulsivity was negatively correlated ($p < .05$) with the duration of the code 'Verification of a position'. Future work could address these factors more profoundly.

In relation to previous work on mental overload as outlined by, i.a., Young and colleagues (2014), the subjective results showed that the observed behaviors on average did not indicate overload as envisioned in Figure 1, because both MWL and EL remained neutral (note that, however, different measures do not always correlate; Hancock & Matthews,

2019; Gerald Matthews et al., 2015). Here, we can at most speak of potentially *impending* overload (i.e., AOI 1 in Figure 1). Future work could explore whether overload could be defined by specific behavior, or by a long sequence and/or duration of behavioral codes. For impending overload, the code 'Freeze' could for example be differentiated into 'decelerating movement' (i.e., automatized motor execution behavior, such as fetching, slowing down) and 'contemplation-style freeze' (i.e., the freeze reflects thinking, without negative affect; cf., potentially our Freeze – Modifier 2). While overload per se could be defined by a 'startle-like freeze' (i.e., a freeze associated with negative affect; cf., potentially our Freeze – Modifier 1 or 3) (cf., freezing as a defensive musculoskeletal response to an abrupt stimulus, related to negative affect in, Lang, Davis, & Öhman, 2000). Differently, nine positioning attempts or a sequence of long-lasting contemplation-style freezes could as well indicate overload. This way, MWL-levels within the upper red zone could be objectively defined as a function of performance - as called for in earlier work - and could even help define, e.g., a 'maximum permissible cognitive load' constraint for assembly work (Grier et al., 2008; S. Hart & Wickens, 2008).

In all, the proposed MWL measure thus showed to be valid within this specific context to measure impending overload (and potentially mental fatigue and startle-like responses) with five behavioral codes. Future work could also explore effects of overload per se to cover the full spectrum of upper red zone MWL and will hereto need to address possible crucial moderating variables such as interpersonal differences in behavioral expressions, personality, intelligence and experience.

6.1 Limitations

The operationalization of some codes depended on the participant's eye-gaze for the annotation of their onset and end, or to define them (e.g., redirecting eye gaze). It should nevertheless be possible to redesign eye gaze out of the coding scheme, as we tested this with a sample dataset. Also using regular stationary camera's instead of the frontal camera of eye-tracking glasses should be well feasible and could even provide richer datasets since more behaviors are captured.

We decided not to focus on (impending) underload, as the behavior observed in our piloting iterations did not show an adequate amount of code occurrences such as sighing or gazing away, while such behaviors can also indicate (impending) overload. Future work could additionally explore this construct, underrepresented in the literature (Young et al., 2014). We also found that 'Freeze' showed low interrater-reliability for two participants. Next, the fact that raters were not blind to the conditions might have allowed for some subjectivity. Future work could address this better.

Finally, the proposed coding scheme only applies for our specific assembly context. We believe the designed codes can however be customized. Future work can hence explore the generalizability and external validity of these codes when applied in other contexts and with operators (instead of a with a student sample), finding out that, e.g., for some contexts some codes will not be indicative, will not occur, will have to be redefined or that new codes arise. In the early iterations of the current coding scheme we for instance also included sighing as previous work linked such utterances to high MWL (see Vlemincx, Diest, & Bergh, 2012; Vlemincx, Taelman, Peuter, & Diest, 2011). Here, we found that our respective code was too unreliable to code manually. Other behaviors that might apply more in other contexts could be; changing the task execution sequence, re-consulting instructions, verbal expressions such as cursing, ironic laughter, thinking out loud, asking for help, or non-verbal utterances as shaking one's head or fidgeting, and more elementary motor execution such as hovering of the hands over parts or a lower velocity of motor execution in general (cf., Stork & Schubö, 2010). The authors therefore look forward to iterations on our method in a diversified set of contexts with a diverse sample of operator profiles, in which the current codes will be customized and new behavioral codes could arise. This future work will eventually show to what extent the presented codes are generic already and easily customizable.

6.2 Implications for research and practice

Our results are well in line with earlier work assessing, for example, computer mouse activity and behavioral entropy. By designing a behavioral coding scheme for MWL detection in manual assembly, we hope to inspire research further developing the method into a measurement system applicable in (real-life industrial) contexts requiring

unobtrusiveness, ease-of-use and a straightforward way to map MWL fluctuations directly onto redundant or even dangerous behavior. A method with such ecological, external and face validity could hence be of added value to practitioners analyzing and (re)designing assemblies, instructions, assistive technology and training methods (cf., Error Management Training; Keith & Frese, 2008). This might also be relevant for situations in which the consequences of hesitative behavior can be detrimental (cf., the petrochemical industry, surgery).

Concurrent and incremental validity of the proposed behavioral coding method of and beyond other in literature proposed MWL measures (i.e. physiological measures) is recommended as well. Research further scrutinizing physiological measures could also benefit when triangulating by narrowing down to coded events, so that the sensitivity of measures (cf., Matthews et al., 2015) could raise significantly. Also, the aim should be to explore and extend the method's 'group-to-individual generalizability' (Fisher, Medaglia, & Jeronimus, 2018). That is, to investigate to what extent the method can be used at the individual level (instead of at the group level as validated here) and how to promote this further. An experimental approach with many repeated measures within the same subjects but over time (instead of cross-sectional) will be needed for this (Fisher et al., 2018).

The practice of coding requires intensive training, is very labor-intensive and is even prone to attentional blindness (i.e. by focusing on certain behaviors, not perceiving other behaviors). Future research could aim to simplify codes, to make codes generic (cf., freezing) or to develop behavioral recognition algorithms applied on video images to automatically detect, e.g., freezes or positioning attempts. Only then our method could reach the level of widespread industrial applicability. Finally, as Virtual Training Systems (implementing virtual environment, virtual reality or augmented reality) becomes increasingly valuable (cf., Al-ahmari, Abidi, & Ahmad, 2016; Langley et al., 2016), a MWL-measure for such environments could be developed by deriving, e.g., freezes or positioning attempts from the gyroscope data of the controllers.

7. Conclusion

The presented work endeavored to develop a non-obtrusive MWL-measure with high ecological and face validity, as current research on physiological measurement faces reliability challenges in applied settings. A behavioral video coding scheme was developed and revealed to be able to detect multiple assembly behaviors indicative of upper red zone MWL. That is, the occurrence and duration of the behavioral codes 'Parts collection' and 'Positioning', and the occurrence of 'Part rotation', 'Positioning attempt' and 'Freeze - Modifier 2' was significantly higher during the execution of a high complexity assembly, as compared to a low complexity assembly. The authors aspire to hereby initiate a novel line of measurement validation to decisively help practitioners and operators optimize MWL-levels on the shop floor.

8. Acknowledgments

This work was supported by the strategic research centre for the manufacturing industry Flanders Make, Oude Diestersebaan, 133, 3920 Lommel, Belgium, as part of the SBO project 'Augmented workers using smart robots in a manufacturing cell (Yves)'. The authors report no conflict of interest.

9. Declaration of interests

The authors declare no conflict of interest.

10. References

Al-ahmari, A. M., Abidi, M. H., & Ahmad, A. (2016). Development of a virtual manufacturing assembly simulation system. *Advances in Mechanical Engineering*, *8*(3), 1–13. https://doi.org/10.1177/1687814016639824

Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review*, *22*(4), 425–438. https://doi.org/10.1007/s10648-010-9130-y

Arico, P., Borghini, G., Di Flumeri, G., Sciaraffa, N., & Babiloni, F. (2018). Passive BCI beyond the lab: Current trends and future directions. *Physiological Measurement*, *39*(8), aad57e. https://doi.org/10.1088/1361-6579/aad57e

Arshad, S., Wang, Y., & Chen, F. (2013). Analysing mouse activity for cognitive load

detection. In *Proceedings of the 25th Australian Computer-Human Interaction Conference on Augmentation, Application, Innovation, Collaboration - OzCHI '13* (pp. 115–118). New York, New York, USA: ACM Press. https://doi.org/10.1145/2541016.2541083

Badarna, M., Shimshoni, I., Luria, G., & Rosenblum, S. (2018). The importance of pen motion pattern groups for semi-automatic classification of handwriting into mental workload classes. *Cognitive Computation*, *10*(2), 215–227. https://doi.org/10.1007/s12559-017-9520-2

Baranek, G. T., Barnett, C. R., Adams, E. M., Wolcott, N. A., Watson, L. R., & Crais, E. R. (2005). Object play in infants with autism: Methodological issues in retrospective video analysis. *American Journal of Occupational Therapy*, *59*(1), 20–30. https://doi.org/10.5014/ajot.59.1.20

Bates, D., Maechler, M., Bolker, B., Walker, S., Bojesen Christensen, R. H., Singmann, H., … Fox, J. (2019). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-21. Retrieved from https://github.com/lme4/lme4/

Boer, E. R. (2000). Behavioral entropy as an index of workload. *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Association, "Ergonomics for the New Millennium,"* 125–128. https://doi.org/10.1177/154193120004401702

Booker, J. D., Swift, K. G., & Brown, N. J. (2005). Designing for assembly quality: Strategies, guidelines and techniques. *Journal of Engineering Design*, *16*(3), 279–295. https://doi.org/10.1080/09544820500126672

Brolin, A., Thorvald, P., & Case, K. (2017). Experimental study of cognitive aspects affecting human performance in manual assembly. *Production and Manufacturing Research*, *5*(1), 141–163. https://doi.org/10.1080/21693277.2017.1374893

Brookhuis, K. A., & de Waard, D. (2001). Assessment of drivers' workload: Performance, subjective and physiological indices. In P. A. Hancock & P. A. Desmond (Eds.), *Stress, workload, and fatigue* (pp. 321–333). Mahwah, NJ: L. Erlbaum.

Brouwer, A. M., Zander, T. O., van Erp, J. B. F., Korteling, J. E., & Bronkhorst, A. W. (2015). Using neurophysiological signals that reflect cognitive or affective state: Six recommendations to avoid common pitfalls. *Frontiers in Neuroscience*, *9*(APR), 1–11. https://doi.org/10.3389/fnins.2015.00136

Bruya, B. (2010). *Effortless attention. A new perspective in the cognitive sciene of attention and action*. Cambridge, MA: MIT Press. A Bradford book.

Cain, B. (2007). A review of the mental workload literature. In *Report No.: RTO-TR-HFM-121-Part-II.* (pp. 1–34). Ontario, WT: Defence Research and Development Canada Toronto Human System Integration Section.

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, *74*, 221–232. https://doi.org/10.1016/j.apergo.2018.08.028

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust multimodal cognitive load measurement*. (D. Tan & J. Vanderdonckt, Eds.), *Human–Computer Interaction Series*. Switzerland: Springer. https://doi.org/10.1007/978-3-319-31700-7_3

Chorney, J. M. L., McMurtry, C. M., Chambers, C. T., & Bakeman, R. (2015). Developing and modifying behavioral coding schemes in pediatric psychology: A practical guide. *Journal of Pediatric Psychology*, *40*(1), 154–164. https://doi.org/10.1093/jpepsy/jsu099

Cohen, J. (1960). A coefficient of agreement for nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

de Waard, D. (1996). *The measurement of drivers' mental workload. PhD dissertation*. University of Groningen, Haren. https://doi.org/10.1016/j.apergo.2003.11.009

de Waard, D., & Lewis-Evans, B. (2014). Self-report scales alone cannot capture mental workload. *Cognition, Technology & Work*, *16*(3), 303–305. https://doi.org/10.1007/s10111-014-0277-z

de Winter, J. C. F. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective. *Cognition, Technology and Work*, *16*(3), 289–297. https://doi.org/10.1007/s10111-014-0275-1

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192–203. https://doi.org/10.1037/1040-3590.18.2.192

ElMaraghy, W., ElMaraghy, H., Tomiyama, T., & Monostori, L. (2012). Complexity in engineering design and manufacturing. *CIRP Annals*, *61*(2), 793–814. https://doi.org/10.1016/j.cirp.2012.05.001

Erol, S., Jäger, A., Hold, P., Ott, K., & Sihn, W. (2016). Tangible industry 4.0: A scenario-based approach to learning for the future of production. *Procedia CIRP*, *54*, 13–18. https://doi.org/10.1016/j.procir.2016.03.162

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Foy, H. J., & Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Applied Ergonomics*, *73*(April), 90–99. https://doi.org/10.1016/j.apergo.2018.06.006

Friard, O., & Gamba, M. (2016). BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution*, *7*(11), 1325–1330. https://doi.org/10.1111/2041-210X.12584

Grier, R., Wickens, C., Kaber, D., Strayer, D., Boehm-Davis, D., Trafton, J. G., & St. John, M. (2008). The red-line of workload: Theory, research, and design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(18), 1204–1208. https://doi.org/10.1177/154193120805201811

Guru, K. A., Shafiei, S. B., Khan, A., Hussein, A. A., Sharif, M., & Esfahani, E. T. (2015). Understanding cognitive performance during robot-assisted surgery. *Urology*, *86*(4), 751–757. https://doi.org/10.1016/j.urology.2015.07.028

Hairston, W. D., Whitaker, K. W., Ries, A. J., Vettel, J. M., Bradford, J. C., Kerick, S. E., & McDowell, K. (2014). Usability of four commercially-oriented EEG systems. *Journal of Neural Engineering*, *11*(4). https://doi.org/10.1088/1741-2560/11/4/046018

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hancock, P. A., & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human Factors*, *61*(3), 374–392. https://doi.org/10.1177/0018720818809590

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload.* (pp. 139–183). Amsterdam: Elsevier Science Publishers B.V. (North-Hollland).

Hart, S., & Wickens, C. D. (2008). Mental Workload. In *NASA Human Integration Design Handbook*.

Heyman, R. E., Lorber, M. F., Eddy, J. M., & West, T. V. (2018). Behavioral observation and coding. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 345–372). New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511996481.018

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive energetical framework. *Biological Psychology*, *45*(96), 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

Hu, S. J., Zhu, X., Wang, H., & Koren, Y. (2008). Product variety and manufacturing complexity in assembly systems and supply chains. *CIRP Annals*, *57*(1), 45–48. https://doi.org/10.1016/j.cirp.2008.03.138

Kahneman, D. (1973). Attention and effort. *The American Journal of Psychology*, *88*(2), 339. https://doi.org/10.2307/1421603

Keith, N., & Frese, M. (2008). Effectiveness of error management training: A meta-analysis. *Journal of Applied Psychology*, *93*(1), 59–69. https://doi.org/10.1037/0021-9010.93.1.59

Lang, P. J., Davis, M., & Öhman, A. (2000). Fear and anxiety: Animal models and human cognitive psychophysiology. *Journal of Affective Disorders*, *61*(3), 137–159. https://doi.org/10.1016/S0165-0327(00)00343-8

Langley, A., Lawson, G., Hermawati, S., D'Cruz, M., Apold, J., Arlt, F., & Mura, K. (2016). Establishing the usability of a virtual training system for assembly operations within the automotive industry. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *26*(6), 667–679. https://doi.org/10.1002/hfm.20406

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2010). Practical resources for assessing and reporting intercoder reliability in content analysis research projects. Retrieved from http://matthewlombard.com/reliability/index_print.html

Longo, F., Nicoletti, L., & Padovano, A. (2017). Smart operators in industry 4.0: A human-

centered approach to enhance operators' capabilities and competencies within the new smart factory context. *Computers and Industrial Engineering*, *113*, 144–159. https://doi.org/10.1016/j.cie.2017.09.016

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Mandler, G. (1979). Thought processes, consciousness, and stress. In V. Hamilton & D. M. Warburton (Eds.), *Human stress and cognition: an information processing approach.* (pp. 179–201). New York: John Wiley & Sons, Inc.

Matthews, G. (2016). Multidimensional profiling of task stress states for human factors: A brief review. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(6), 801–813. https://doi.org/10.1177/0018720816653688

Matthews, G., De Winter, J., & Hancock, P. A. (2019). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science*, 1–31. https://doi.org/10.1080/1463922X.2018.1547459

Matthews, Gerald, Reinerman-Jones, L. E., Barber, D. J., & Abich, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human Factors*, *57*(1), 125–143. https://doi.org/10.1177/0018720814539505

Matthews, Gerald, Szalma, J., Rose, A., Neubauer, C., & Warm, J. S. (2013). Profiling task stress with the dundee stress state questionnaire. In L. Cavalcanti & S. Azevedo (Eds.), *Psychology of stress: new research* (pp. 49–91). Hauppage, NY: Nova Science Publishers.

McAlister, A. M., Lee, D. M., Ehlert, K. M., Kajfez, R. L., Faber, C. J., & Kennedy, M. S. (2017). Qualitative coding: An approach to assess inter-rater reliability. *ASEE Annual Conference and Exposition, Conference Proceedings*, *2017-June*. https://doi.org/10.18260/1-2--28777

McHugh, M. L. (2012). Lessons in biostatistics interrater reliability: The kappa statistic. *Biochemica Medica*, *22*(3), 276–282. Retrieved from https://hrcak.srce.hr/89395

McKendrick, R., Parasuraman, R., Murtza, R., Formwalt, A., Baccus, W., Paczynski, M., & Ayaz, H. (2016). Into the wild: Neuroergonomic differentiation of hand-held and augmented reality wearable displays during outdoor navigation with functional near infrared spectroscopy. *Frontiers in Human Neuroscience*, *10*.

https://doi.org/10.3389/fnhum.2016.00216

Meulenbroek, R. G. J., Van Galen, G. P., Hulstijn, M., Hulstijn, W., & Bloemsaat, G. (2005). Muscular co-contraction covaries with task load to control the flow of motion in fine motor tasks. *Biological Psychology*, *68*(3), 331–352. https://doi.org/10.1016/j.biopsycho.2004.06.002

Montani, F., Vandenberghe, C., Khedhaouria, A., & Courcy, F. (2020). Examining the inverted U-shaped relationship between workload and innovative work behavior: The role of work engagement and mindfulness. *Human Relations*, *73*(1), 59–93. https://doi.org/10.1177/0018726718819055

Morton, J., Vanneste, P., Larmuseau, C., Acker, B. B. Van, Raes, A., Bombeke, K., … De, L. (2019). Identifying predictive EEG features for cognitive overload detection in assembly workers in Industry 4.0. In *3rd International Symposium on Human Mental Workload: Models and Applications (H-WORKLOAD 2019)*. Rome.

Neuendorf, K. A. (2002). *The Content Analysis Guidebook.* Thousand Oaks, CA: Sage.

Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (1st ed., pp. 7–15). New York: Lawrence Erlbaum Associates Inc.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*, 44–64.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Paas, F. G. W. C., van Merriënboer, J. J. G., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, *79*(1), 419–430. https://doi.org/10.2466/pms.1994.79.1.419

Paras, C. R., Yang, S., Tippey, K., & Ferris, T. K. (2015). Physiological indicators of the cognitive redline. *Proceedings of the Human Factors and Ergonomics Society*, *2015-Janua*, 637–641. https://doi.org/10.1177/1541931215591139

Parmentier, D. D., Van Acker, B. B., Detand, J., & Saldien, J. (2019). Design for assembly meaning: A framework for designers to design products that support operator cognition during the assembly process. *Cognition, Technology & Work*. https://doi.org/10.1007/s10111-019-00588-x

Pillay, H. K. (1997). Cognitive load and assembly tasks: Effect of instructional formats on

learning assembly procedures. *Educational Psychology*, *17*(3), 285–299. https://doi.org/10.1080/0144341970170304

Qiu, J., & Helbig, R. (2012). Body posture as an indicator of workload in mental work. *Human Factors*, *54*(4), 626–635. https://doi.org/10.1177/0018720812437275

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 185–218). Amsterdam: North-Holland: Elsevier Science Publishers B.V.

Richardson, M., Jones, G., & Torrance, M. (2004). Identifying the task variables that influence perceived object assembly complexity. *Ergonomics*, *47*(9), 945–964. https://doi.org/10.1080/00140130410001686339

Richardson, M., Jones, G., Torrance, M., & Baguley, T. (2006). Identifying the task variables that predict object assembly difficulty. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(3), 511–525. https://doi.org/10.1518/001872006778606868

Rosenthal, R. (1994). Parametric measures of effect size. In C. H & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York, NY: Russell Sage Foundation.

Ruiz, N., Taib, R., & Chen, F. (2006). Examining the redundancy of multimodal input. In *Proceedings of the 20th conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction: design: activities, artefacts and environments - OZCHI '06* (p. 389). New York, New York, USA: ACM Press. https://doi.org/10.1145/1228175.1228254

Shalin, V. L., Prabhu, G. V., & Helander, M. G. (1996). A cognitive perspective on manual assembly. *Ergonomics*, *39*(1), 108–127. https://doi.org/10.1080/00140139608964438

Snijders, T., & Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Applied Multilevel Analysis*. SAGE Publications Ltd, London.

Srygley, J. M., Mirelman, A., Herman, T., Giladi, N., & Hausdorff, J. M. (2009). When does walking alter thinking? Age and task associated findings. *Brain Research*, *1253*, 92–99. https://doi.org/10.1016/j.brainres.2008.11.067.When

Staal, M. A. (2004). Stress, cognition, and human performance: A literature review and conceptual framework. *NASA Technical Memorandum 212824*, pp 1-170.

Stinissen, J. (1977). Revised minnesota paper form board test. Vorm AB - Leuvense aanpassing van vorm MA en MB, Likert, R. & Quasha, W. (pp. 1–7). Amsterdam: Swets & Zeitlinger, B.V.

Stork, S., & Schubö, A. (2010). Human cognition in manual assembly: Theories and applications. *Advanced Engineering Informatics*, *24*(3), 320–328. https://doi.org/10.1016/j.aei.2010.05.010

Um, J., Lyons, A., Lam, H. K. S., Cheng, T. C. E., & Dominguez-Pery, C. (2017). Product variety management and supply chain performance: A capability perspective on their relationships and competitiveness implications. *International Journal of Production Economics*, *187*, 15–26. https://doi.org/10.1016/j.ijpe.2017.02.005

Van Acker, B. B., Bombeke, K., Durnez, W., Parmentier, D. D., Mateus, J. C., Biondi, A., … Vlerick, P. (2020). Mobile pupillometry in manual assembly: A pilot study exploring the wearability and external validity of a renowned mental workload lab measure. *International Journal of Industrial Ergonomics*, *75*, 102891. https://doi.org/10.1016/J.ERGON.2019.102891

Van Acker, B. B., Parmentier, D. D., Vlerick, P., & Saldien, J. (2018). Understanding mental workload: From a clarifying concept analysis toward an implementable framework. *Cognition, Technology and Work*, *20*(3). https://doi.org/10.1007/s10111-018-0481-3

Vanneste, P., Raes, A., Morton, J., Bombeke, K., Van Acker, B. B., Larmuseau, C., … Van den Noortgate, W. (2020). Measuring cognitive load during assembly work through multimodal physiological data. *Unpublished Manuscript*.

Vera, D., Crossan, M., Rerup, C., & Werner, S. (2014). "Thinking before acting" or "acting before thinking": Antecedents of individual action propensity in work situations. *Journal of Management Studies*, *51*(4), 603–633. https://doi.org/10.1111/joms.12075

Vitense, H. S., Jacko, J. A., & Emery, V. K. (2003). Multimodal feedback: An assessment of performance and mental workload. *Ergonomics*, *46*(1–3), 68–87. https://doi.org/10.1080/00140130303534

Vlemincx, E., Diest, I. Van, & Bergh, O. Van Den. (2012). A sigh following sustained attention and mental stress: Effects on respiratory variability. *Physiology & Behavior*, *107*, 1–6. https://doi.org/10.1016/j.physbeh.2012.05.013

Vlemincx, E., Taelman, J., Peuter, S. D. E., & Diest, I. V. A. N. (2011). Sigh rate and respiratory variability during mental load and sustained attention. *Psychophysiology*,

48, 117–120. https://doi.org/10.1111/j.1469-8986.2010.01043.x

Wan, X., & Sanders, N. R. (2017). The negative impact of product variety: Forecast bias, inventory levels, and the role of vertical integration. *International Journal of Production Economics*, *186*, 123–131. https://doi.org/10.1016/j.ijpe.2017.02.002

Wickens, C. D. (2002). Multiple resources and performance prediction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *3*(2), 159–177.

Wickens, Christopher D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2014). State of science: Mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17. https://doi.org/10.1080/00140139.2014.956151

11. Appendices

Appendix A : Overview of coding instructions

***Category*, Code (Unit)**	**Definition**	**Coding Instructions**
I. Selecting Behavior		
1. Parts collection (state event)	Time needed to select all the perceived correct parts.	From first glance on parts, for at least 2 seconds, until onset positioning (starting at first glance for at least two seconds on the assembly-in-progress).
2. Part touched (point event)	Number of new parts that are touched.	Code until onset positioning. Nested in 'Parts collection'.
3. Part rotation (point event)	Number of times the operator is manipulating a part (here rotating, from velocity is 0 to velocity is 0).	Also rotating part back to original position counts as rotation. Code until onset positioning. Nested in 'Parts collection'.
II. Positioning Behavior		
4. Positioning (state event)	Time to select perceived right position. Includes rotation and alignment.	From eye gaze switch from selected part (defined as part that will be positioned) to assembly-in-progress fixating on assembly-in-progress for at least 2s, until onset first screw being inserted or eye-gaze at not-yet selected parts in case a new 'Parts collection' occurs.
5. Positioning attempt (point event)	Number of positions tried (defined by velocity = 0, part touches assembly-in-progress or is being held above, below or aside from it) on assembly-in-progress.	Positions already tried count as a new attempt if part has been moved from position. Parts selected in previous steps count as well. Code until the perceived right position (defined by the final position that precedes completely screwing a part) is achieved of the part on the assembly-in-progress. Nested in 'Positioning'.
6. Verification of fixed	Time needed for the operator to verify if a	From finishing screwing (one of the screws of a part) until the next assembly behavior, such as inserting the next screw, or

position (state event)	position (i.e., orientation, alignment, position) is correct after a component has been fastened, for at least 2 seconds.	until eye-gaze for at least 2 seconds on next part to be collected or positioned, until onset next step (here from onset turning around towards laptop with instructions), until checking steadiness of fastening, etc. Not nested in 'Positioning'.
7. Correction of perceived wrong position (state event)	Time the operator needs to correct a (set of) perceived wrong fastening(s) and perform a perceived correction, for at least 2 seconds.	From onset completely unfastening the perceived wrong fixation(s) until onset of screwing the last component of a perceived correction, for at least 2s. Not nested in 'Positioning.'
III. Transition Behavior		
8. Inspection of assembly-in-progress (state event)	Time the operator needs to familiarize with or verify the assembly-in-progress including touching and rotating assembly-in-progress, for at least 2 seconds, in between visual intake of instructions and onset 'Parts collection' or 'Correction of perceived wrong position'.	Can only occur after a first step has been performed, since from then on an assembly-in-progress is present. Code from start gaze position on assembly-in-progress for at least 2 seconds, until start 'Parts collection'.
IV. Generic Behavior		
9. Freeze (state event)	Number of times both operator's hands and arms stop moving or only move very minimally (the latter not as a function of selection or positioning), at a fixed place in the work space for at least 2 seconds.	From start movement stops (minimally) until onset movement.
Modifier 1	Head freezes, eyes move	
Modifier 2	No freeze of eyes and head	
Modifier 3	Both eyes and head freeze too	
10. Gaze redirection (point event)	Number of times operator gazes away from work table for at least 2s.	Code until switching back to eye-gaze on work table for at least 2 seconds.
11. Relative head position change (point event)	Number of times operator's head is repositioned closer to or further away from parts or assembly-in-progress, because of the operator moving the upper body or oppositely, the parts or assembly-in-progress.	Moving back from coded position does not count as new code (e.g., moving closer to the parts or assembly-in-progress while moving back to original position 10s later on, counts as one point event).

Appendix B : Overview of attention points

***Category*, Code (Unit)**	**Attention points**
I. Selecting Behavior	
1. Parts collection (state event)	1. Can also alternate with 'Positioning', in case of 'Positioning' already started but participant gazes back at parts not-yet touched (i.e., parts still on display at outer side of work table). In this case code from first glance on parts until first next glance on assembly-in-progress for at least 2s. 'Parts collection' can hence also only include visual scanning.
2. Part touched (point event)	1. The assembly-in-progress does not count as a part, except during the first step in which there is no assembly-in-progress yet. 2. Parts already touched hence do not count. 3. Can also occur during 'Positioning' or 'Fastening'.
3. Part rotation (point event)	1. Exceptions: rotating assembly-in-progress (this is included in Positioning).
II. Positioning Behavior	
4. Positioning (state event)	1. Can alternate with 'Parts collection' (reminder: if eye-gaze fixates on assembly-in-progress or at not-yet selected parts for at least 2 seconds). 2. In case operator positions two parts at once: code stops at inserting first screw of first part and next 'Positioning' code starts from switch of gaze on second part to at least two seconds on assembly-in-progress and stops at inserting first screw of the second part. 3. For the first step of the assembly, the assembly-in-progress is defined as the part on which the other two parts are being positioned on. 4. Can also occur with parts selected in previous steps.
5. Positioning attempt (point event)	/
6. Verification of fixed position (state event)	1. Includes twisting the component just screwed for alignment and orientation. 2. Checking screws and components for tightness does not count, even if component is being twisted in function of this checkup. 3. Includes visually checking the component in relation to the other components (already fastened). 4. Visually checking other components on the assembly-in-progress for at least 2s without gazing back at original component counts as 'Positioning' instead. 4. Manipulating other components after the respective component has been screwed does not count, but instead counts as a new 'Positioning' code.
7. Correction of perceived wrong position (state event)	1. Can also occur at the onset of a step, with parts assembled in previous steps. 2. Can alternate with 'Parts collection' and 'Positioning' (of newly selected parts). In case component is put aside to complete later on, code should be stopped at point of putting aside and starts again when picking up this component.
III. Transition Behavior	
8. Inspection of assembly-in-progress (state event)	/
IV. Generic Behavior	
9. Freeze (state event)	/
Modifier 1	/
Modifier 2	/
Modifier 3	/
10. Gaze redirection (point event)	/

11. Relative head position change (point event)	/

Appendix C : Overview of results per behavioral code

1. Parts Collection - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		-0.018			-0.018
Dexterity		0.12			0.12
Condition			-0.786***	-1.347***	-1.347***
Step 2			-0.206	-0.786**	-0.786**
Step 3			-0.187	-0.449	-0.449
Condition x Step 2				1.160**	1.160**
Condition x Step 3				0.524	0.524
Constant	0	0	0.524**	0.804***	0.804***
Observations	144	144	144	144	144
R^2	0.02	0.02	0.24	0.31	0.31

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

1. Parts Collection - Duration					
	I	II	III	IV	V
Spatial Intelligence		0.014			0.014
Dexterity		0.101			0.101
Condition			-0.908***	-1.465***	-1.465***
Step 2			-0.863***	-1.235***	-1.235***
Step 3			-1.099***	-1.562***	-1.562***
Condition x Step 2				0.745*	0.745*
Condition x Step 3				0.926**	0.926**
Constant	0	0	1.108***	1.386***	1.386***
Observations	144	144	144	144	144
R^2	0.00	0.01	0.43	0.47	0.48

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

2. Part Touched - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		0.123			0.123
Dexterity		0.160*			0.160*

Condition			-0.308*	-0.482	-0.482
Step 2			-0.793***	-0.944***	-0.944***
Step 3			-1.014***	-1.124***	-1.124***
Condition x Step 2				0.301	0.301
Condition x Step 3				0.221	0.221
Constant	0	0	0.756***	0.843***	0.843***
Observations	144	144	144	144	144
R^2	0.00	0.03	0.24	0.25	0.25

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

3. Part rotation - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		0.02			0.02
Dexterity		0.033			0.033
Condition			-0.498***	-0.614*	-0.614*
Step 2			-0.988***	-1.054***	-1.054***
Step 3			-1.094***	-1.201***	-1.201***
Condition x Step 2				0.133	0.133
Condition x Step 3				0.214	0.214
Constant	0	0	0.943***	1.001***	1.001***
Observations	144	144	144	144	144
R^2	0.00	0.00	0.31	0.31	0.31

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

4. Positioning - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		0.056			0.056
Dexterity		0.018			0.018
Condition			-0.911***	-1.236***	-1.236***
Step 2			0.325	-0.13	-0.13
Step 3			0.228	0.195	0.195
Condition x Step 2				0.911*	0.911*
Condition x Step 3				0.065	0.065

Constant	0	0	0.271	0.434*	0.434*
Observations	144	144	144	144	144
R^2	0.00	0.00	0.23	0.27	0.27

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

4. Positioning - Duration					
	I	II	III	IV	V
Spatial Intelligence		-0.025			-0.025
Dexterity		0.038			0.038
Condition			-1.348***	-1.114***	-1.114***
Step2			-0.009	-0.094	-0.094
Step3			0.286*	0.721***	0.721***
Condition x Step 2				0.17	0.17
Condition x Step 3				-0.872**	-0.872**
Constant	0	0	0.582***	0.465**	0.465**
Observations	144	144	144	144	144
R^2	0.00	0.00	0.48	0.55	0.54

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

5. Positioning Attempt - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		0.006			0.006
Dexterity		0.009			0.009
Condition			-1.068***	-1.222***	-1.222***
Step 2			-0.267	-0.524*	-0.524*
Step 3			-0.121	-0.096	-0.096
Condition x Step 2				0.512	0.512
Condition x Step 3				-0.051	-0.051
Constant	0	0	0.663***	0.740***	0.740***
Observations	144	144	144	144	144
R^2	0.00	0.00	0.30	0.32	0.32

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

6. Verification of position - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		-0.003			-0.003
Dexterity		0.003			0.003
Condition			-0.149	-0.448*	-0.448*
Step 2			0.168	-0.056	-0.056
Step 3			-0.560**	-0.784**	-0.784**
Condition x Step 2				0.448	0.448
Condition x Step 3				0.448	0.448
Constant	0	0	0.205	0.355	0.355
Observations	144	144	144	144	144
R^2	0.09	0.09	0.22	0.24	0.24

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

6.Verification of position - Duration					
	I	II	III	IV	V
Spatial Intelligence		-0.045			-0.045
Dexterity		0.092			0.092
Condition			-0.481**	-0.323	-0.323
Step 2			0.101	0.409	0.409
Step 3			-0.467*	-0.538*	-0.538*
Condition x Step 2				-0.616	-0.616
Condition x Step 3				0.141	0.141
Constant	0	0	0.363*	0.284	0.284
Observations	144	144	144	144	144
R^2	0.12	0.11	0.27	0.30	0.30

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

7. Correction of wrong position - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		0.057			0.057
Dexterity		0.053			0.053
Condition			-0.678***	-0.062	-0.062

Step 2			0.586**	1.233***	1.233***
Step 3			0.216	0.493	0.493
Condition x Step 2				-1.295***	-1.295***
Condition x Step 3				-0.555	-0.555
Constant	0	0	0.072	-0.236	-0.236
Observations	144	144	144	144	144
R^2	0.00	0.00	0.18	0.28	0.27

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

7. Correction of wrong position - Duration					
	I	II	III	IV	V
Spatial Intelligence		0.054			0.054
Dexterity		0.190*			0.190*
Condition			-0.470**	0.026	0.026
Step 2			0.359	0.763**	0.763**
Step 3			0.398*	0.738**	0.738**
Condition x Step 2				-0.808*	-0.808*
Condition x Step 3				-0.680*	-0.680*
Constant	0	0	-0.017	-0.265	-0.265
Observations	144	144	144	144	144
R^2	0.10	0.08	0.21	0.25	0.24

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

8. Inspection of assembly-in-progress - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		-0.009			-0.009
Dexterity		0.113*			0.113*
Condition			-0.188*	0.042	0.042
Step 3			0.104	0.333**	0.333**
Condition x Step 3				-0.458**	-0.458**
Constant	0.281**	-0.068	0.323***	0.208*	-0.141
Observations	96	96	96	96	96

R^2	0.00	0.08	0.07	0.17	0.20

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 2 were used as reference categories.

8. Inspection of assembly-in-progress - Duration					
	I	II	III	IV	V
Spatial Intelligence		-0.077			-0.077
Dexterity		0.878			0.878
Condition			-1.851*	-0.704	-0.704
Step 3			0.706	1.853	1.853
Condition x Step 3				-2.295	-2.295
Constant	1.917***	-0.665	2.489***	1.916*	-0.666
Observations	96	96	96	96	96
R^2	0.00	0.06	0.06	0.08	0.14

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 2 were used as reference categories.

9. Freeze (Modifier 1) - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		-0.099			-0.099
Dexterity		0.011			0.011
Condition			-0.802***	-0.463	-0.463
Step 2			0.116	0.324	0.324
Step 3			0.069	0.37	0.37
Condition x Step 2				-0.417	-0.417
Condition x Step 3				-0.602	-0.602
Constant	0	0	0.339*	0.17	0.17
Observations	144	144	144	144	144
R^2	0.00	0.01	0.16	0.18	0.19

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

9. Freeze (Modifier 1) - Duration					
	I	II	III	IV	V
Spatial Intelligence		-0.034			-0.034
Dexterity		0.071			0.071

Condition			-0.764***	-0.278	-0.278
Step 2			0.205	0.491	0.491*
Step 3			0.243	0.688*	0.688*
Condition x Step 2				-0.571	-0.571
Condition x Step 3				-0.889*	-0.889*
Constant	0	0	0.233	-0.01	-0.01
Observations	144	144	144	144	144
R^2	0.00	0.01	0.16	0.19	0.20

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

9.Freeze (Modifier 2) - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		0.033			0.033
Dexterity		-0.004			-0.004
Condition			-0.673***	-0.561*	-0.561*
Step 2			0.056	0.112	0.112
Step 3			0.112	0.224	0.224
Condition x Step 2				-0.112	-0.112
Condition x Step 3				-0.224	-0.224
Constant	0	0	0.280	0.224	0.224
Observations	144	144	144	144	144
R^2	0.01	0.01	0.17	0.17	0.17

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

9. Freeze (Modifier 2) - Duration					
	I	II	III	IV	V
Spatial Intelligence		-0.053			-0.053
Dexterity		-0.012			-0.012
Condition			-0.574***	-0.451	-0.451
Step 2			0.037	0.074	0.074
Step 3			0.147	0.294	0.294
Condition x Step 2				-0.074	-0.074
Condition x Step 3				-0.294	-0.294

Constant	0	0	0.226	0.164	0.164
Observations	144	144	144	144	144
R^2	0.00	0.00	0.10	0.11	0.10

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

10. Gaze redirection - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		-0.024			-0.024
Dexterity		0.184*			0.184*
Condition			-0.107	-0.322	-0.322
Step 2			0.081	-0.161	-0.161
Step 3			0.081	0	0
Condition x Step 2				0.484	0.484
Condition x Step 3				0.161	0.161
Constant	0	0	0	0.107	0.107
Observations	144	144	144	144	144
R^2	0.07	0.04	0.07	0.09	0.06

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.

11. Relative head position change - Occurrence					
	I	II	III	IV	V
Spatial Intelligence		-0.026			-0.026
Dexterity		0.018			0.018
Condition			-0.316	-0.526	-0.526
Step 2			-0.053	-0.105	-0.105
Step 3			-0.263	-0.526	-0.526
Condition x Step 2				0.105	0.105
Condition x Step 3				0.526	0.526
Constant	0	0	0.263	0.368	0.368
Observations	144	144	144	144	144
R^2	0.00	0.00	0.04	0.05	0.05

The values in the table are standardized weights (Beta's).
*p<0.05; **p<0.01; ***p<0.001
N = 24
Low spatial intelligence, low dexterity, high complexity and Step 1 were used as reference categories.