Wandahl, S., Pérez, C. T., Salling, S., &Lerche, J. (2022). Robustness of Work Sampling for Measuring Time Waste. *Proceedings of the 30th Annual Conference of the International Group for Lean Construction (IGLC30)*, 247–258. doi.org/10.24928/2022/0127

ROBUSTNESS OF WORK SAMPLING FOR MEASURING TIME WASTE

Søren Wandahl[1], Cristina Toca Pérez[2], Stephanie Salling[3], Jon Lerche[4]

ABSTRACT

Construction can be considered a socio-technical system, which is challenging to model due to the many agents interacting either in a managed way or autonomously. Therefore, cause and effect models are hard to validate, and a traditional correlation approach is insufficient. In this study, the method of robustness testing was applied to test the effect stability when assumptions of a model are changed. The research objective is to apply robustness testing on WS data to assess the robustness and validity of the WS method. An actual refurbishment project was the case for this study, where data was acquired through nine days of continuous WS application. Time-series data were grouped into Direct Work (DW), Indirect Work, and Waste Work. Several different robustness tests were applied. It can be concluded that the WS method is robust, i.e., the effect (DW) is stable even if the assumptions are changed severely. Deleting 90% of the sample does, for instance, almost not change the effect. Likewise, if errors are infused into the sample, the effect is stable. Also, if certain structural parts are excluded from the sample, e.g., observations during morning startup, etc., the effect is still stable.

KEYWORDS

Value stream, Waste, Trust, Robustness, Work Sampling

INTRODUCTION

Construction is often described as a complex project system (Bertelsen, 2003; Lindhard & Wandahl, 2013). The concept of why and how a project is complex has developed over time. Williams (1999) describes two dimensions of complexity. Firstly, structural complexity (Baccarini, 1996) is the number of elements in a system and the interdependence of the elements. Elements can be both organizational and product-wise. Secondly, the degree of uncertainty in both how well defined the project's goals are and how well defined the methods of achieving those goals are. Later, three additional dimensions of complexity were added to the understanding (Geraldi et al., 2011). The first, dynamics, refers to changes in projects, i.e., changes in specifications. Changes are enforced on a project from both outside and inside. Changes lead to rework, disorder, and

[1] Professor, Department of Civil & Architectural Engineering, Aarhus University, Denmark, swa@cae.au.dk, https://orcid.org/0000-0001-8708-6035

[2] Postdoc, Department of Civil & Architectural Engineering, Aarhus University, Denmark, cristina.toca.perez@cae.au.dk, https://orcid.org/0000-0002-4182-1492

[3] Research Assistant, Department of Civil & Architectural Engineering, Aarhus University, Denmark, stsa@cae.au.dk, https://orcid.org/0000-0001-7088-6458

[4] Postdoc, Department of Business Development and Technology, Aarhus University, Denmark, jon.lerche@btech.au.dk, https://orcid.org/0000-0001-7076-9630

inefficiency. The second, pace, is a type of complexity, as urgency and criticality of time and goals require managerial attention. The third, socio-technical complexity, is supported by a strong stream of research that stresses that projects are carried out by human actors with potentially conflicting interests and incompatible personalities.

All of the abovementioned dimensions of complexity are often present in large construction projects, sometimes resulting in poor performance. Both an effort to analyze root causes of low performance and an effort to improve performance by infusing new innovations, structures, procedures, etc., depend on the rationale that nothing happens without reason, i.e., effects have causes. Due to the complexity of construction, it is difficult to use a simple correlation of one effect based on one cause. Thus, it is very hard in construction to prove a causal relationship of performance and cause.

Nonetheless, academics often try to develop different models of construction that attempt to show how a complex socio-technical system like construction works. The purpose of a model is usually to visualize, understand, or optimize. They can range from simple models with few variables and components (e.g., input-output model or a black box diagram) to larger and more complex models with many variables and components (e.g., Building Information Modeling (BIM) including time). The beauty of a model is that it is not reality; it is a simplification. This fundamental understanding of the abstraction is frequently forgotten or misunderstood, as some researchers and practitioners tend to think that a model is a one-to-one representation of the real world. Many models are either misinformed. i.e., contain errors, wrong assumptions, etc., or are under-informed, i.e., too little data and information levels are too low. It can rightfully be assumed that most models are misinformed or under-informed; thus, they are challenged on their validity (Neumayer & Plümper, 2017).

Accepting that models are only a simplified representation of a social-technical system gives rise to the importance of assessing a model’s validity. Determining the strength of a correlation of two variables as a means of validity for a cause and effect is insufficient given the complex nature of construction projects. Instead, robustness can be introduced to determine how valid a model of a social-technical system, like construction, is. Robustness is a way of assessing the effect stability of a model when assumptions and structures of the model are gradually changed (Neumayer & Plümper, 2017).

The objective of this research is to devise a method for assessing the robustness and validity of the WS method.

The following part of the paper describes the theoretical background in two parts. First, Work Sampling as a way of measuring and modeling time waste in construction. Second, Social Complexity and Robustness as a method of assessing a model’s validity.

MODELING TIME WASTE IN CONSTRUCTION

One of the areas, the Lean Construction community has struggled to model, is measuring time waste. Questions like ‘how can time waste be measured?’, ‘what are the root causes of time waste?’, and ‘how does implementing Last Planner and other Lean approaches reduce time waste?’ are addressed in several research studies, e.g., (Bølviken & Kalsaas, 2011; Kalsaas et al., 2014; Lerche et al., 2022; Neve et al., 2020; Wandahl et al., 2021).

Bølviken & Kalsaas (2011) recognized a need for a more valid method for measuring time waste. Thus, they reviewed a number of direct and indirect measurement methods, and Kalsaas (2011) concluded on the method selection that a suitable method for measuring workflow should mainly be based on the Work Sampling (WS) method.

WORK SAMPLING TO MEASURE TIME WASTE

The WS method has been used for decades to collect data on the amount of value-adding work time, referred to as Direct Work (DW) in the WS method (Gong et al., 2011). WS is a quantitative method applying direct observations to obtain data on how workers use their time on the construction site. In general, WS has been applied throughout time to improve, often single construction projects regarding efficiency, construction labor productivity, and construction cost and time. Thomas (1981) provides relevant insights on how a WS study can be planned and how the collected data can be analyzed. In this research, the present authors apply a more statistical approach to WS in order to validate the robustness of the method in general. However, the authors still acknowledge that WS should mainly be applied to improve a single construction project.

The WS method quantifies how much time workers use on direct work and other categories of preparatory work and waste work. All WS studies apply a DW category. However, the picture is more blurred when it comes to the preparatory and waste work categories. Some studies categorize all none direct work time as waste, while other studies have a more detailed view of non-value-adding work. Generally speaking, non-value-adding work time can in WS be divided into Indirect Work (IW) and Waste Work (WW), resulting in WS having three categories of time; DW, IW, and WW. Work Sampling and, in particular, the share of DW's relation to productivity has been debated throughout time, as DW directly influences the denominator and indirectly the numerator of the productivity equation. However, recent studies conclude that DW is statistically significantly correlated to construction labor productivity at activity, project, and national levels (Araujo et al., 2020; Neve et al., 2020; Siriwardana et al., 2017) and, thus, can be applied as an acceptable indicator for productivity.

CRITIQUE OF WORK SAMPLING

Wandahl et al. (2021) identified 474 case studies where WS was applied in construction. Thus, it can be concluded that the method is widely used. Nonetheless, a severe critique of the method exists, e.g., Josephson & Björkman (2013). Several of the critical points are related to the robustness of the WS method and the potential lack of causality.

Categorizing work activities into direct work and subcategories of preparatory and waste work is very inconsistent (Josephson & Björkman, 2013; Wandahl et al., 2021). This makes cross-case comparison difficult, like any longitudinal meta-analysis (Horman & Kenley, 2005; Josephson & Björkman, 2013). However, it seems that the consequence of inconsistent categorization has not been further researched. In relation to the categorization, Johansen et al. (2021) discovered that, in particular, preparatory work is often considered as direct work by many practitioners and also by some academics. This despite that Ohno (1988) clearly articulated which kinds of activities are value-adding and which are not. The inconsistent understanding of value-adding and non-value-adding work has also led to a critique of WS relying on individual observers (Jenkins and Orth, 2004; Josephson & Björkman, 2013). These observers might be biased and have a non-aligned understanding of waste and value (Neve et al., 2020).

CAUSAL COMPLEXITY AND ROBUSTNESS

When developing a model based on empirical data, it is an interpretation of the actual phenomenon. To capture the true processes of a complex world, researchers would need to precisely know the set of regressors, include all relevant variables and exclude all irrelevant variables, operationalize and measure these variables correctly, etc. (Neumayer & Plümper, 2017). This is not possible. Researchers today agree that a model cannot be specified correctly due to causal complexity. Traditionally, the strategy is to apply assumptions and to accept underdetermined models. The aim of an underdetermined model is a simplified model. However, underdetermination often ends in misspecification, as it requires intensive knowledge to simplify in a valid way (Neumayer & Plümper, 2017). The misspecification of models is a well-known problem, and as Box & Draper (1987) concluded: "*All models are wrong, but some are useful.*" Therefore, researchers must find the optimal trade-off between simplicity and generality to ensure models are not misspecified, as misspecified models lead to biased conclusions.

Causal complexity is an extension of the causality concept, which is often related to correlation. Causality is the study of how things influence one another and how causes lead to effects. There are a number of basic assumptions in a classical (and physics-related) understanding of causation. Firstly, things (effects) have causes. They do not just happen of their own accord. Secondly, effects follow causes in a predictable, linear manner. E.g., concrete cures faster if you increase the ambient temperature. Thirdly, big effects can grow from several small causes, e.g., several minor variations in activity durations can suddenly cause a delay of an entire construction project. Having identified a cause-effect relationship, it often becomes relevant to measure the strength of this relationship. As elaborated later in this paper, it can be difficult to precisely express and measure the strength of such a cause-effect relationship. Often, statistical measures are applied to consider the relationship between two variables, a course variable (the predictor variable) and the effect variable (the response variable). This is referred to as the statistical correlation of cause and effect. However, correlation does not always imply causation. It is two different measures that can, however, be coinciding. In the world of classical physics, this is often the case, and correlation can be a good indicator of, e.g., the causation between ambient temperature and concrete maturity.

CAUSAL COMPLEXITY

Causal complexity is the interpretation of cause-effect in social science. It differentiates from the classical understanding of causation on five important dimensions (Neumayer & Plümper, 2017): (1) Cause-effect relationships in the social world are probabilistic instead of deterministic, therefore, the probability of an effect is a continuum from 0 (a cause does not have an effect) to 1 (the cause is deterministic); (2) Causal complexity is the existence of conditional causal effects and heterogeneous causal effects. Some causes only have an effect if certain conditions are satisfied (Franzese, 2003); (3) The timing of cause and effect. Scholars all too often implicitly assume that an effect occurs immediately after a cause. Yet, effects can occur with a delayed onset; (4) In the real world, effects can precede causes. Human beings have rational expectations about potential future effects and may already act on their expectations rather than on the cause itself. This is called the Cause-precedes-law; and (5) Effects can affect non-treated causes. In the social world, spill-over effects from the treated to the untreated are likely.

ROBUSTNESS AND ROBUSTNESS TESTING

Acknowledging the existence of causal complexity as the boundary conditions for causes and effects on construction sites, another measure than the traditional correlation assessment is needed to assess the strength of a relationship between variables in a model built to simulate construction. Instead, the model's robustness must be tested by systematically removing or changing the assumptions. However, it is difficult to give an unequivocal definition of robustness, as this concept is differently defined in several domains. When investigating the different application domains of robustness, the lack of a unique definition becomes visible. In project management, Robust Decision Making is defined as "*a set of concepts, processes, and enabling tools that use computation, not to make better predictions, but to yield better decisions under conditions of deep uncertainty*" (Lempert, 2019). This is similar to the definition of robustness in statistics, which is "*Robust statistics addresses the problem of making estimates that are insensitive to small changes in the basic assumptions of the statistical models employed*" (Fabozzi et al., 2014). Insensitivity is also the core of robustness in scheduling, as "*a schedule is robust if its performance is rather insensitive to the data uncertainties [...]*" (Billaut et al., 2008). In this research, robustness is defined in a simple way as "effect stability." In WS, this equals measuring DW and $CI_{95\%}$ stability when assumptions and sample size are altered.

When acknowledging construction sites as a phenomenon in which causal complexity exists, a need to investigate any model of that phenomenon for robustness arises. The robustness testing is to test and analyze the uncertainty of a model and test whether estimated effects of interest are sensitive to changes in model specifications. The literature describes an extensive range of robustness grouped into model variation, permutation, limit, and placebo tests. The following method section describes which robustness tests were applied in this research.

RESEARCH METHOD

The WS method was used in the case study to collect a data set that could be used for robustness testing. The case consisted of a social housing refurbishment project of 24 five-story buildings. The main renovation tasks were related to carpentry work, such as replacing windows and roofs and installing new ventilation and electrical systems.

WS data were collected during nine days, named Day 1 to Day 9, with observations from work begins in the morning until work ends in the afternoon (i.e., from 07.00-15.30, excluding break times from 09.00-09.15 and 11.30-12.00). Two different observers, named Observer A and B, randomly toured the construction site. The WS method was applied in seven trades, constituting 40 workers. After completing the nine days of data collection, 1,550 random observations (representing a sample of N=1,550) were recorded with a 95% confidence interval ($CI_{95\%}$) of ± 3.42%. In order to avoid patterns of behavior and to reduce the variability of the measurement, the authors collected a homogeneous sample. The average of the sample was 172 observations per day, with a standard deviation of around 43 observations, through the nine days of data collection (see the table in Figure 1), resulting in around three observations every 15 minutes. In this study, a six-work category classification was adopted. The applied categorization follows the method of Activity Analysis (CII, 2010), which outlines which work activities must be put into which categories.The six categories are: (1) production, e.g., installing gypsum boards; (2) talking, e.g., discussing the installation process; (3) preparation, e.g., measuring with a ruler; (4) transportation, e.g., carrying tools; (5) walking, e.g., moving empty-handed; and (6) waiting, e.g., delaying action until receiving material.

ROBUSTNESS TESTING

After data collection was completed, robustness testing was applied. Robustness testing was conducted in four steps, according to the approach outlined by Neumayer & Plümper (2017): (1) Define the subjectively optimal specification for the data-generating process at hand, i.e., the baseline model; (2) Identify assumptions made in the baseline model, which are potentially arbitrary; (3) Develop models that change one of the baseline model's assumptions at a time; and (4) Compare the estimated effects of each robustness test model to the baseline model and compute the estimated degree of robustness.

Step 1: Defining the baseline model

The first step consisted of defining the baseline model. In this case, the baseline model was the actual WS data collected, consisting of the 1,550 random observations collected. Figure 1 shows the results of the baseline, including a stabilization curve, 95% confidence interval, split between the WS categories, and information on the data collection.

	N	DW	IW	WW
Day 1 (Tue)	210	25%	51%	24%
Day 2 (Wed)	245	15%	49%	36%
Day 3 (Thu)	161	30%	60%	9%
Day 4 (Fri)	110	34%	55%	11%
Day 5 (Mon)	207	24%	62%	14%
Day 6 (Tue)	156	19%	58%	23%
Day 7 (Wed)	184	32%	45%	23%
Day 8 (Thu)	152	20%	51%	29%
Day 9 (Fri)	125	30%	46%	24%
Total	**1550**	**25%**	**53%**	**22%**

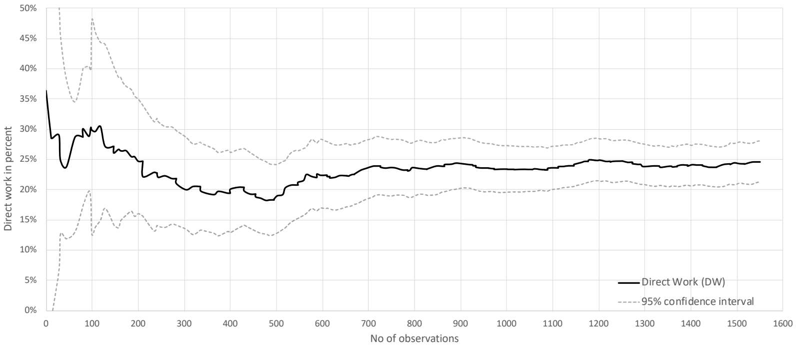

Figure 1: Baseline of work sampling data.

Step 2: Defining assumptions in the model

The second step, defining assumptions in the WS model, was a brainstorming session to identify important assumptions. Five fundamental assumptions in the WS model were identified (i) Each workday is similar, i.e., observations are uniform; (ii) Direct Work stabilizes after around 550 observations; (iii) Productive and preparatory work can be distinguished based on momentary observations; (iv) A few observation errors do not influence the overall result; and (v) Results are independent of the observers.

Step 3: Defining Robustness test models

Three different types of robustness tests were applied: (1) Model variation; (2) Randomized permutation; and (3) Structured permutation. In the Model variation tests, assumptions (i) & (ii) are tested. In the randomized permutation tests, assumptions (iii) & (iv) are tested. In the final structured permutation test, assumptions (i) & (v) are tested.

Firstly, the model variation tests change one, or sometimes more, model specification assumptions and replace them with an alternative assumption. Our analysis changed the sample size, both reversibly, by deleting data points from the end of the data collection period towards the beginning, and randomly from 0% to 100%. A Monte Carlo Simulation of 500 simulations was conducted for each alteration to analyze the effect stability (change in DW and 95% Confidence Interval, $CI_{95\%}$).

Secondly, a randomized permutation test was conducted on different assumptions. Random permutation tests change specification assumptions repeatedly. Errors were infused randomly into the sample to monitor effect stability in one analysis. Again, 500 runs of Monte Carlo Simulation were conducted. An error is a faulty observation, i.e., interpreted or noted into a wrong category in the WS study. It is common that production, preparation, and talking get confused and wrongly noted. This was analyzed, applying Monte Carlo Simulation to investigate the effect stability.

Thirdly, a structured permutation test was conducted on the specific assumption in the WS model. Structured permutation tests change a model assumption within a model space systematically. Changes in the assumption are based on a rule rather than random. Different structures, i.e., parts of the sample, were deliberately excluded in the Monte Carlo simulation, like specific days, the first hour, observation after lunch, observer 1, observer 2, etc. Again, the effect stability on DW and $CI_{95\%}$ were observed.

Step 4: Robustness testing analysis

The fourth and final step, comparing results to the baseline, was conducted to discuss and interpret the results. Lastly, the authors presented some of the main implications for practitioners of the present analysis.

FINDINGS – INTERPRETING THE ROBUSTNESS TESTING

MODEL VARIATION TESTS

The model variation tests investigated the effect of stability when changing the sample size. The first test was to reduce the sample size reversibly, starting from N=1,550. The effect is illustrated on the stabilization graph, cf. figure 1. DW is stable from N=700, which is after day 4. In other words, reducing the sample size by 55% did not influence DW or the 95% confidence interval. Another approach was to reverse the sample size until DW exceeded the final 95%-confidence interval. At N=1,550 DW is 24.65% and $CI_{95\%}$ is ±3.42%. The sample size was, thus, reversed until it exceeded 24.65%±3.42%, which occurred at N=559 (after day 3), where the lower confidence interval was exceeded.

A second model variation test reduced the sample size randomly. A random reduction is an irreversible reduction of the sample size. Table 1 illustrates the effect stability of DW and $CI_{95\%}$ corresponding to a random deletion of observation, i.e., a random reduction of sample size. Results are based on 500 Monte Carlo simulations.

Table 1: Random deletion of observations resulting in a random decrease in sample size.

Sample	N	DW	$CI_{95\%}$
Baseline (N=100%)	1,550	24.65%	3.42%
Random (N=90%)	1,395	24.54%	3.49%
Random (N=80%)	1,240	24.61%	3.62%
Random (N=70%)	1,085	24.51%	3.74%
Random (N=60%)	930	24.57%	3.91%
Random (N=50%)	775	24.56%	4.18%
Random (N=40%)	620	24.47%	4.48%
Random (N=30%)	465	24.61%	5.01%
Random (N=20%)	310	24.45%	5.81%
Random (N=10%)	155	24.78%	7.51%

Table 1 shows that a random decrease in sample size had almost no influence on DW but clearly increased the $CI_{95\%}$ interval, making the data less valid. Nonetheless, reducing the sample size by 50% only increased the $CI_{95\%}$ by 22.22%.

RANDOMIZED PERMUTATION TESTS

The first randomized permutation test investigated assumption (iv) by looking at the effect stability if the observer made mistakes. There are two types of mistakes;

misinterpreting an observation and assigning an observation to the wrong trade or work samling category. 500 Monte Carlo simulations were conducted for each change, cf table 2. Table 2 shows that randomly changing categories affected both DW and $CI_{95\%}$, however, the impact was insignificant. 20% error equals 310 errors or 4.6 errors per observed hour, which impacted DW with 10.5%.

Table 2: Random error in categories.

Sample	N	DW	$CI_{95\%}$
Baseline	1,550	24.65%	3.42%
5% error, Random category	1,550	23.92%	3.37%
10% error, Random category	1,550	23.39%	3.32%
20% error, Random category	1,550	22.07%	3.21%
30% error, Random category	1,550	20.85%	3.10%
40% error, Random category	1,550	19.58%	3.02%

The second random permutation test was more realistic, as it is not likely that the observer mistakes, e.g., walking for production and so on. Realistically, the observer could misinterpret preparation with production and vice versa, and talking with production and vice versa. The effect stability of such confusion is shown in table 3. Once again, the results in table 3 were based on 500 Monte Carlo simulations.

Table 3: Production to preparation and vice versa.

Change	Baseline	5%	10%	15%	20%	25%
	DW±$CI_{95\%}$	DW±$CI_{95\%}$	DW±$CI_{95\%}$	DW±$CI_{95\%}$	DW±$CI_{95\%}$	DW±$CI_{95\%}$
Preparation to production	24.65% ±3.42%	25.87% ±3.47%	27.11% ±3.51%	28.13% ±3.55%	29.47% ±3.60%	30.72% ±3.64%
Talking to production	24.65% ±3.42%	24.80% ±3.42%	25.41% ±3.43%	25.96% ±3.44%	26.32% ±3.45%	26.50% ±3.45%
Production to talking or prepa.	24.65% ±3.42%	23.30% ±3.31%	22.11% ±3.21%	20.71% ±3.09%	19.79% ±3.01%	18.39% ±2.88%

Table 3 shows that a change from the value-adding DW category to the preparatory category of Indirect Work or vice-versa influences the DW. Johansen et al. (2021) concluded that the observer misinterpreting preparation as production is the most common error. If 25% of the preparation observations are misinterpreted or wrongly assigned to production, DW is 30.72%, which almost equals one-third of the work time. One-third of the work time being productive is often referred to as state-of-the-art.

STRUCTURED PERMUTATION TEST

A structured permutation test infuses structured and logical changes in the baseline model. Table 4 shows that most of the structured exclusions had a limited impact on the stability.

In this case, assumptions regarding the uniformity of observation days and time of the day, and the independence of the observer were analyzed. The sensitivity of DW and $CI_{95\%}$ were analyzed based on excluding designated parts of the sample as described above. Only excluding the first, the last, or both the first and last hour of the day had an impact on the DW stability higher than 1 percent point. Most significant is the result when observer A or B is excluded. This has a significant impact on the average DW and the confidence interval.

Table 4: Exclusion of structured part of the Work Sampling.

Sample size	N	DW	$CI_{95\%}$
Baseline	1,550	24.65%	3.42%
Excluding mornings (from 07.00-11.00)	676	24.70%	5.08%
Excluding afternoons (from 11.30-15.30)	874	24.60%	4.66%
Excluding first hour (from 07.00-08.00)	1,310	26.95%	3.76%
Excluding last hour (from 14.30-15.30)	1,443	25.71%	3.61%
Excluding both the first and last hour	1,203	28.43%	3.99%
Excluding observer A	398	25.88%	8.26%
Excluding observer B	697	27.40%	4.32%
Excluding Mondays	1,343	24.72%	3.76%
Excluding Tuesdays	1,184	25.34%	4.11%
Excluding Wednesdays	1,121	25.60%	3.99%
Excluding Thursday	1,237	24.41%	3.68%
Excluding Fridays	1,315	23.35%	3.62%

DISCUSSION

In WS, there is an assumption that the share of DW is an effect of efficient management and planning. However, there is no single cause variable, as multiple factors will affect the DW share. Therefore, WS does not fit well with the traditional concept of causality. Josephson & Björkman (2013) argues that WS can not be used for cross-case comparisson as there are too many factors influencing the share of DW. In other words, a single cause-effect relationhsip can not be devised based on WS, which is a limitation of the method. This research confirms that limitation. WS and DW as a response variable should instead be understood in the light of causal complexity.

The five dimensions in causal complexity suit well with WS. Improved management and planning have a probabilistic impact on DW and cannot be modeled with 100% precision. In addition, the effects are heterogeneous and depend on an unknown mix of conditions. WS is time-sensitive because one cannot expect the effect (improved DW share) right after implementing new management and planning initiatives. It might be delayed, and it might fluctuate. As construction is a social-technical system with many actors, improved DW might be measurable without any causes implemented, merely due to the expectation of effects among workers. Moreover, there is a likely spill-over as one trade with optimized planning and management can improve DW of other trades that have not received implementation. Based on these five dimensions, causal complexity can be used to understand and reject some of the critique of work sampling that has been raised based on a traditional correlation and cause-effect thinking.

The robustness testing of WS also can reject some part the WS critique, i.e. lack of causality. However, the critique rasied that the misinterpretation of VAW and NVAW will influence the WS result still remains. Also the dependence of the observer was raised as a critue, and this has also not been possible to reject based on the robustness testing.

IMPLICATIONS FOR PRACTITIONERS

The results provided a new angle to the body of knowledge for WS by utilizing the robustness method to understand WS measures, contributing to the ongoing discussion of productivity in both construction (Neve et al., 2020) and offshore wind (Lerche et al., 2022). However, it also raises a question regarding the levels of productivity that today are considered state-of-art. In particular, if DW is not adequately separated from IW and WW. As the random permutation test in table 3 showed, an incorrect categorization of an observation (production vs. preparation and talking) will have a direct impact on DW. That is, 10% faulty registration will result in a change in DW of 10%. Therefore, the WS method is still considered sensitive toward the categorization of observations.

On the other hand, the structured permutation test showed that the WS is robust towards structural changes in the observation patterns. Most structural changes in observation patterns only had a limited effect on DW. However, excluding the first and last hour of the day did have some impact. This can be explained by the start and stop of the day, where less production is going on, as time is spent on preparing, moving, cleaning, etc. This is in line with Neve et al. (2020), who concluded that inparticular starts and stops are critical to gaining high labor productivity.

From a practical perspective, the findings show how misinterpretation of work categories can transform less promising results into state-of-art results. Meanwhile, the robustness testing also revealed that random sampling, even with fewer observations, can still be considered significant and provide a proper indication of productivity. Therefore, the methodology can easily be applied by practitioners without being too worried about the potential faulty application.

CONCLUSION

This research aimed to apply robustness testing on a WS data set gathered in a real construction project to assess the robustness and validity of the WS method. That objective has successfully been achieved.

This paper discussed that the widespread assumption considering that the share of time spent on DW in the construction process is an effect of efficient management and planning cannot be explained considering a single cause variable, as multiple factors will affect the DW share. Because of that, the robustness test can be considered a suitable method to test and analyze the uncertainty of the work sampling method.

After analyzing the data collected in a single case study, it can be concluded that the WS method is robust. Three different types of robustness testing were conducted, and most changes in assumptions, sample size, structure, and internal logic in the WS method only had a limited effect on the average DW result and its confidence interval. Most of the critique of WS cited in this work can, thus, be refuted. Only the dependence of the observer and the categorization of DW and preparatory work need attention and require more research in the future to conclude upon finally.

ACKNOWLEDGMENTS

This work was supported by the Independt Research Fund Denmark (grant no. 0217-00020B). The authors are also grateful for the construction company that opened its door for this study.

REFERENCES

Araujo, L. O. C., Neto, N. R., and Caldas, C. H. (2020). “Analyzing the Correlation between Productivity Metrics.” *Construction Research Congress 2020*.

Baccarini, D. (1996). “The concept of project complexity - a review.” *International Journal of Project Management*, 14, 201-204.

Bertelsen, S. (2003). “Construction as a Complex System.” *Proc., The 11th annual conference in the International Group for Lean Construction*.

Billaut, J., Moukrim, A., and Sanlaville, E. (2008). *Flexibility and Robustness in Scheduling* Willey.

Bølviken, T., and Kalsaas, B. T. (2011). “Discussion of Strategies for Measuring Workflow in Construction.” *Proc., 19th Annual Conference of the International Group for Lean Construction*Lima, Peru.

Box, G. E. P., and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*, Wileys, New York.

CII (2010). “A Guide to Activity Analysis.” *Construction Industry Institute,* 2010

Fabozzi, F. J., Focardi, S. M., Rachev, S. T., and Arshanapalli, B. G. (2014). *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications.*, John Wiley & Sons.

Geraldi, J. Maylor, H., and Williams, T. (2011). “Now, let's make it really complex (complicated).” *International Journal of Operations & Production Management*, 31 (9), 966-990

Gong, J., Borcherding, J. D., and Caldas, C. H. (2011). “Effectiveness of craft time utilization in construction projects.” *Construction Management and Economics*, 29(July), 737-751.

Horman, M. J., and Kenley, R. (2005). “Quantifying Levels of Wasted Time in Construction with Meta-Analysis.” *Journal of Construction Engineering and Management*, 131(1), 52-61.

Jenkins, J. L., and Orth, D. L. (2004). “Productivity improvement through work sampling.” *Cost Engineering*, 46(3), 27-32.

Johansen, P., Christensen, S., Neve, H. H., and Wandahl, S. (2021). “Lean Renovation – a Case Study of Productivity, Flow, and Time Improvements.” *Proc., Proc. 29th Annual Conference of the International Group for Lean Construction (IGLC)*Lima, Peru, 839-848.

Josephson, P., and Björkman, L. (2013). “Why do work sampling studies in construction? The case of plumbing work in Scandinavia.” *Engineering, Construction and Architectural Management*, 20(6), 589-603.

Kalsaas, B. T. (2011). “On the Discourse of Measuring Work Flow Efficiency in Construction. A Detailed Work Sampling Method.” *Proc., 19th Annual Conference of the International Group for Lean Construction*Lima, Peru.

Kalsaas, B. T., Gundersen, M., and Berge, T. O. (2014). “To measure workflow and waste. A concept for continuous improvement.” *Proc., IGLC22*.

Lempert, R. J. (2019). "Robust Decision Making (RDM).” *Decision Making under Deep Uncertainty: From Theory to Practice”*, V. A. W. J. Marchau, W. E. Walker, P. J. T. M. Bloemen, and S. W. Popper, eds., Springer International Publishing, Cham, 23-51.

Lerche, J., Lindhard, S., Enevoldsen, P., Neve, H. H., Møller, D. E., Jacobsen, E. L., Teizer, J., and Wandahl, S. (2022). “Causes of delay in offshore wind turbine construction projects.” *Production Planning & Control*, 1-14.

Lerche, J., Lorentzen, S., Enevoldsen, P., and Neve, H. H. (2022). "The impact of COVID -19 on offshore wind project productivity – A case study." *Renewable and Sustainable Energy Reviews*, 158, 112188.

Lindhard, S., and Wandahl, S. (2013). "Scheduling of Large, Complex, and Constrained Construction Projects. An Exploration of LPS Application." *International Journal of Project Organisation and Management*, 6(3), 237-253.

Neumayer, E., and Plümper, T. (2017). *Robustenss tests for Quantitative Research*, Cambridge University Press.

Neve, H. H., Wandahl, S., Lindhard, S., Teizer, J., and Lerche, J. (2020). "Determining the Relationship between Direct Work and Construction Labor Productivity in North America: Four Decades of Insights." *Journal of Construction Engineering and Management*, 146(9), 04020110.

Neve, H. H., Wandahl, S., Lindhard, S., Teizer, J., and Lerche, J. (2020). "Learning to see value-adding and non-value-adding work time in renovation production systems." *Production Planning and Control*.

Ohno, T. (1988). *Toyota Production System*, Productivity Press, New York.

Siriwardana, C. S. A., Titov, R., and Ruwanpura, J. (2017). "A study to investigate the relationships between work sampling categories and productivity for concrete slab formwork." *Proc., 8th International Conference on Structural Engineering and Construction Management*.

Thomas, H. R. (1981). "Can work sampling lower construction costs." *Journal of Construction Division*, 107(2), 263-278.

Wandahl, S., Neve, H. H., and Lerche, J. (2021). "What a Waste of Time." *Proc., Proc. 29th Annual Conference of the International Group for Lean Construction (IGLC)*Lima, Peru, 157-166.

Williams, T.M. (1999). "The need for new paradigms for complex projects." *International Journal of Project Management*, 17(5), 269-273.