# Evaluating the Impact of an Assembly System 4.0 on Human Error

Jacqueline Humphries[1,*], Pepijn Van de Ven[2], and Alan Ryan[3]

[1]Department of Information Technology, Applied Science Technology Faculty, Technological University of the Shannon, Limerick, Ireland
[2]Department of Information Technology, Science & Engineering Faculty, University of Limerick, Limerick, Ireland
[3]Department of Engineering, Science & Engineering Faculty, University of Limerick, Limerick, Ireland
Email: jacqueline.humphries@tus.ie (J.H.); pepijn.vandeven@ul.ie (P.V.d.V.); alan.ryan@ul.ie (A.R.)
*Corresponding author

Manuscript received December 25, 2023; revised January 20, 2024; accepted February 3, 2024; published April 16, 2024

***Abstract*—Assembly Systems assist the human in manufacturing by providing easy access to the assembly procedure. However, they are static systems that provide the same information to all users. Assembly Systems 4.0 are a relatively new concept that use data-driven insights to assist the human by providing context-specific information about the assembly process. Assembly Systems 4.0 should have a positive impact on reducing human error in manufacturing. In this research the utility of an Assembly System 4.0 is evaluated. Two experiments are conducted to investigate potential effects on human performance from the lens of error and failed quality parts. Through the first experiment, a laboratory simulation, it is proven that the Assembly System 4.0 can detect human error. In the second experiment, the system is compared against a traditional Assembly System. Four hundred assemblies are conducted, in a two-independent sample test. It is found that neither system prevents human error from occurring. However, how the error is treated is significant. The users of the Assembly System 4.0 detected, and corrected the errors, producing higher quality products.**

***Keywords*—assembly, quality, error, production**

## I. Introduction

Assembly Systems (AS) support the human in an assembly process by providing access to the Standard Operating Procedures (SOP). Typically, this is done by displaying the assembly instructions on a screen in text, drawing and/or photographic form. The human follows the assembly instructions, and confirms that that step is complete, upon which the next step in the assembly process is presented. The cycle continues until the whole of the manufacturing process is complete.

Presenting the assembly instructions at the work station through AS, supports the human to achieve high quality assembly outputs [1]. Despite the widespread availability of AS, however, error in assembly continues to be a major problem for manufacturing companies [2]. Thus discovering new ways to reduce the likelihood of human error in manufacturing by AS is an important research gap to address. The digital capacity available under Industry 4.0 can be applied in new ways to assembly systems with the aim of supporting better data-driven assembly. Such systems are referred to as Assembly Systems 4.0 (AS4.0) [3, 4]. There is a need for empirical evidence as to impact these systems have on improving the work of the human. The objective of this paper is to evaluate the utility of an AS4.0. The hypotheses explored in this study are:

- H1. The use of an AS4.0 will detect errors made in assembly.
- H2. The use of an AS4.0 will reduce the amount of human errors made in assembly.
- H3. The use of an AS4.0 will result in lower rates of failed quality products.

The rest of this paper is structured as follows: first the case study is outlined. In the case study a description of an exemplar AS4.0 is provided, followed by a description of the product to be assembled in the experiments. Next, the research method outlines two experiments that are conducted, with the results and discussion on the same. Finally, potential research gaps and limitations are summarized in the conclusions.

## II. Materials and Methods

Two experiments are conducted. The first is a laboratory simulation in which assembly errors are introduced to provide evidence of the utility and quality of AS4.0 artefacts in detecting error. The second is an experimental simulation to provide empirical evidence of AS4.0 from an operational impact, that is does the human make an error, and does the human rework the error to produce a high quality product whilst using the AS4.0.

### *A. Case Study: A Data-Driven Cyber-Physical Assembly 4.0 System*

In this research, an exemplar Assembly System 4.0 is used as the case study for the hypotheses testing. The system is a fully working complete AS4.0 system, which is installed onto a work station. A simple assembly product is chosen for the experiments. In this section, the system is first described, then the experiments outlined.

The human is presented at the workstation with a touchscreen, and a login page into the system. The human logs in and is verified by the system. The human scans a work order by a digital scan of the work order barcode. The associated production assembly instructions are retrieved. An example workstation is illustrated in Fig. 1. The instructions are provided to the human via a number of interfaces, dependent upon the preferences of the human Fig. 1(a). The human follows the instructions provided to them. An example screen showing video instructions is illustrated in Fig. 2.

High definition cameras monitor the assembly station, Fig. 1(b). The cameras are positioned at different angles and so capture different perspectives of the product being assembled. Images of the assembly are captured and processed by algorithms to detect whether the work is correct or not. Feedback is presented to the human, see Fig. 1(c). If errors are identified by the computer vision algorithms, the human's attention is brought to the same, allowing them the opportunity to rework the assembly so that the error does not get propagated through the system. Several algorithms are used to perform the analysis in a hybrid multi-stage model. Computer vision uses the You Only Look Once model, version 8 (YOLOv8), a Convolutional Neural Network (CNN), for object detection, classification, and segmentation tasks. Classification robustness is tested and false predictions are eliminated by using Inception version 3, an image recognition model. Further, multiple cameras view the object being assembled from different angles. The multiple predictions, determined upon the image feed from each camera, are integrated to produce consistent and accurate information in a process known as data fusion.


Fig. 1. Assembly 4.0 work station.

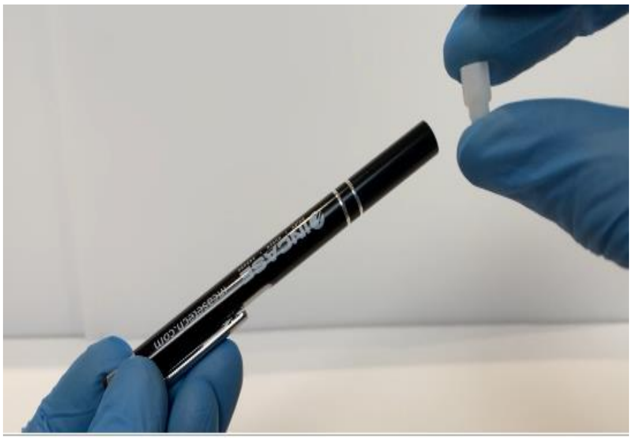

Fig. 2. Close up of the assembly instructions on the screen.

## B. *Product Assembly Use Case*

The use-case chosen is assembly of a pen. The pen is chosen so as to remove any bias with regard to task complexity. Further the pen, being a common product, does not introduce the concerns of the impact the "learning effect" has on error as raised in Stockinger, Polanski-Schräder and Subtil [5] and Riedel *et al.* [6]. The pen assembly comprises of four steps: insert the thrust device the right way up into the barrel, insert the ink chamber into the barrel, add the spring to the ballpoint tip of the ink chamber, and screw the tip onto the barrel.

## C. *Experiment I: Evaluation of System's Ability to Detect Error*

The first experiment addresses the first hypothesis H1: *The use of an AS4.0 will detect errors made in assembly*. A laboratory controlled experiment is chosen to evaluate the AS4.0's utility in this regard. The rationale for the design of this experiment arises from the low numbers of human errors made in assembly normally, which may distort results in an experiment with human participants.

One hundred ($N$ = 100) experiments are conducted by the researcher. The researcher creates errors which are identified by the system. In half of the experiments no error is made, and in the other half error is introduced. There are several different types of error that can be made [7–9], with Error by Omissions (OM), Error by Confusion (CX) such as incorrect selection of similar parts, and Execution Error (EX) such as incorrect selection of system and incorrect fixation, being the most commonly reported [7]. In this experiment, the following errors are evaluated (see left most column of Table 1) which could be expected in a real production environment on the assembly of this device. How error is introduced into the experiment for each type of error, is shown in the right most column of Table 1.

Table 1. Error types and treatment

| **Potential Error Investigated** | **How Error introduced in experiment** |
|---|---|
| Omissions such as process steps not conducted | Insertion of the thrust device omitted from the assembly |
| Incorrect selection of variants | Similar but incorrect parts placed into the totes included different shaped and/or coloured pen tips, ink chambers of the incorrect length, and thrust devices of the incorrect length and shape |
| Incorrect fixation/adjustment | The pen tip is not tightened enough onto the barrel |
| Picked up poor quality parts | Bent and extended springs introduced |
| Wrong counting | The work order required five completed pens, too few and too many are created |

A sample of the correct parts and their corresponding incorrect part used are illustrated in Fig. 3.

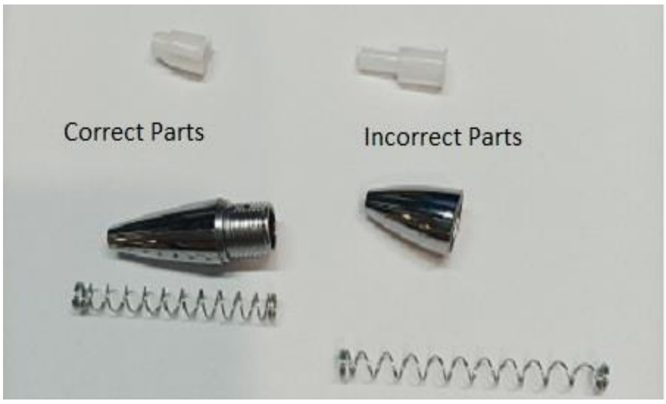


Fig. 3. Sample of correct and incorrect parts.

The assembly is conducted using the assistance from the system. A sample of the data captured is illustrated in Table 2. In test numbers 1 and 16 errors are made in the assembly. For test 1, the system correctly identified the error as a “failed quality”. The boolean data in column “Predict Correct?” column is an evaluation of the accuracy of the system. In the final column “Type of Result” the four types of error states can be accounted for: True Negative (TN), True Positive (TP), False Positive (FP), False Negative (FN).

In this experiment, a “positive” state is the presence of an error, and a “negative” state is that no error occurs. A True Positive occurs when the error is made, and is identified as an error by the system, that is the assembly is wrong and is correctly labelled as wrong. A True Negative occurs when no error is made, and no error is identified by the system, that is the assembly is correct, and is correctly labelled. A False Positive occurs where no error is made, but the system predicts an error, that is the assembly is correct but is incorrectly labelled as wrong by the system. These are Type I errors. A False Negative occurs when an error has been made, but the system predicts there is no error, that is the assembly is wrong but incorrectly labelled as correct. These are Type II errors.

Table 2. Experiment I data

| Test No. | Test Type | Expected Result | Received Result | Predict Correct? | Type of Result |
|---|---|---|---|---|---|
| 1 | OM Left out thrust | FAIL | FAIL | TRUE | True Positive: Error Found |
| 16 | CX Wrong spring Inserted | FAIL | PASS | FALSE | False Negative: Incorrectly missed the error |
| 50 | Insert thrust | PASS | PASS | TRUE | True Negative: No Error Found |

### *D. Experiment II: Evaluation of Impact on Error by Comparison with AS*

The second experiment is an experimental simulation chosen to evaluate the impact of the system on production. Experimental simulation allows for the maximum potential for precise measures of behaviour of human participants. In order to control the environment, the simulation is held in an artificially recreated setting—a research laboratory in the Technological University of the Shannon. A work station, similar to that used in a factory is built, and the system installed.

For a desired power of 0.80 and alpha of 0.5, the required sample size per group is 20 [10]. As the research is interested in error, and based on a pilot test, an error every 10 observations is expected, 200 samples per group are required. Twenty participants took part in the study, who are recruited from the student body. The students had an interest in manufacturing solutions and task design. The participants then conduct 5 assemblies, of 4 steps in each, to create 200 observations per group. Two treatments are used. The first is the Control Treatment (CT) in which the traditional AS is used. The second is the System Treatment (ST) in which the AS4.0 is used. One group of participants are assigned to CT, and one group assigned to ST. The participants are randomly allocated between the treatments, although care is taken to remove bias and so the gender and age profile between the groups is similar. Also the same training and instructions, and the same feedback modes are provided to all participants.

In CT the participants are given the SOP for a manufacturing process through a traditional AS interface and observed completing the production. A simple AS is built that displays a full screen webpage for each step in the process. The instructions are displayed in text and image format on the page. The participant follows the instructions on the page, and presses a button on the touchscreen interface to indicate that that step is complete. The next instruction is loaded, and the process is completed until the operator stops. There is no intelligence built into the AS. It purely presents the instructions for the particular SOP being built on a step-by-step basis. In ST the participants complete production by following the support given by the AS4.0 described in the case study. The machine learning algorithms assess whether the assembly is correct, and provide feedback to the participant. Once the step is complete the instructions for the next step are presented to the participant.

The natural behaviour of the participants is observed and measured. Incorrect selection does cause quality defects, and for this reason, and in order to provide an opportunity for error, the same number of incorrect and correct parts are added to the totes holding the parts prior to the assembly.

The analysis of data from this experiment involves testing multiple null hypotheses simultaneously, as there are multiple outcomes of interest. The multiple hypotheses tested are *H2: The use of an AS4.0 will reduce the amount of human errors made in assembly; H3: The use of an AS4.0 will result in lower rates of failed quality products.* With respect to the hypotheses around error, analysis focuses on comparison of the number of errors within each group. The outcomes of interest are the number of errors made, the errors corrected and the errors remaining in the products after assembly. The hypothesis test is for the difference between medians ($\mu_{1/2}$) of the two populations—CT and ST. The null hypothesis is that the use of AS4.0 has no positive effect on the median human errors made, and remaining after possible correction.

$H2_0$: $\mu_{1/2}$ ErrorsMade_AS-$\mu_{1/2}$ ErrorsMade_AS4.0 = 0

$H2_a$: $\mu_{1/2}$ ErrorsMade_AS-$\mu_{1/2}$ ErrorsMade_AS4.0 > 0

$H3_0$: $\mu_{1/2}$ ErrorsRemain_AS-$\mu_{1/2}$ ErrorsRemain_AS4.0 = 0

$H3_a$: $\mu_{1/2}$ ErrorsRemain_AS-$\mu_{1/2}$ Errors Remain _AS4.0 > 0

To quantify the benefits of the system, the parameters regarding quality as recommended in Keller *et al.* [11] are used and defined as follows:

$$\text{errorRate } (Q_e) = X_{ea} / X_{et}$$

$$\text{errorCorrectionRate } (Q_{ec}) = X_{ec} / X_{ea}$$

$$\text{errorDetectionRate } (Q_{ed}) = X_{ed} / X_{eb}$$

where $X_{ea}$ is the number of errors after the observed process; $X_{et}$ the total number of different errors made by all participants; $X_{ec}$ the number of corrected errors; $X_{ed}$ the number of detected errors; and $X_{eb}$ the number of errors before the observed process has started.

The data is generated by the system and the researcher. Data is recorded, and saved to a spreadsheet file, and cleaned. This data is then read into a data sink, JMP Pro 17, for analysis. Sample data is illustrated in Table 3.

Table 3. Experiment II sample data

| Treatment (CT: Control \| ST: System) | ST |
|---|---|
| Observation No. | 140 |
| Product Assembly No. | 5 |
| Assembly Step | 4 |
| Start Timestamp for Step ($t_{Start}$140.ST) | 07/11/2023 11:15 |
| End Timestamp for Step ($t_{End}$140.ST) | 07/11/2023 11:22 |
| Time Spent on Consultation in seconds | 2 |
| Time Spent on Execution in seconds | 5 |
| Complete Step Time in seconds | 7 |
| Incorrect selection of variants (wrong thrust/nib) | 1 |
| Omissions such as process steps not conducted | 0 |
| Incorrect fixation/adjustment | 0 |
| Picked up poor quality parts | 0 |
| Wrong counting | 0 |
| Errors Made | 1 |
| Error Detected | 1 |
| Error Corrected | 1 |
| Errors Remaining | 0 |
| Error Rate | 0.03 |
| Error Correction Rate | 1.00 |
| Error Detection Rate | 1.00 |
| Work Completed Correctly 1= Correct, 0= Wrong | 1 |

## III. Result and Discussion

The results of the experiments are presented in two sections corresponding to the two experiments: quantitative evaluation of accuracy, and quantitative evaluation of operational benefits with participants.

### *A. Results Experiment I: Evaluation of System's Ability to Detect Error*

In the first experiment, one hundred assemblies are conducted in a controlled environment. An example of the errors as identified by the AS4.0 are illustrated in Fig. 4.


Fig. 4. Detection of wrong variant on thrust selection, and the corrected action by the operator as detected by the system.

The images show the live feed to the cameras, the classification label and a confidence score. Once the confidence score passes a threshold value the assembly in the image is classified as correct or wrong. The threshold values for each step in the assembly are tuned for optimum prediction, with thresholds in the region of 0.60 to 0.82. For Example, in Fig. 4 an incorrect part is picked up, identified as "wrong", and then the correct part picked up, with the action identified as "corrected". The assembly will not be 'passed' by the system until the confidence index passes the corresponding threshold value.

In the confusion matrix provided in Table 4, it can be seen that True Positives (TP) are recorded in 92% of tests where error is made (n: 46/50). The computer vision algorithms correctly identified that the assembly process had not been performed correctly. True Negative (TN) was correctly identified in 98% of all tests taken where error was made (n: 9/50). One False Positive (FP) was identified (n: 1/50). On this occasion (n: 1/50) there was a delay in the computer vision algorithm identifying the assembly. On this particular occasion the computer vision algorithm had missed the thrust device being inserted, as hands were blocking the part from being seen by the camera. The system did not progress to the next step of the assembly. The thrust device had to be removed, and reinserted in a way that the camera could see it. The system then passed the assembly as being correct. False Negatives (FN) were found in 8% of all tests taken where error is made (n: 4/50). In all of these instances the spring or ink chamber was the offending part. The computer vision model only identified the part, not the length of the part. So an ink chamber or a spring protruding a few millimeters further than it should was not identified.

Table 4. Confusion matrix

| | | Actual | |
|---|---|---|---|
| | | Positive 1 | Negative 0 |
| Predict | Positive 1 | TP 46 | FP 1 |
| | Negative 0 | FN 4 | TN 49 |

### *B. Results Experiment II: Evaluation of Comparison of AS4.0 with AS*

Four hundred assemblies ($N$ = 400) are conducted, by twenty participants allocated to two treatments ST and CT ($n_{ST}$ = 200, $n_{CT}$ = 200). Two researchers observed the assemblies and noted error, responses, timings, and cross-referenced the data. A level of significance of $p = 0.05$ is used in all the experiments.

#### *1) Operational benefits with respect to error*

In four hundred assemblies, sixty-three errors are made ($N$ = 63/400). On four occasions more than one error is made on a particular assembly step, so fifty-nine assemblies failed quality checks ($N$ = 59/400). The median error rate and mode error rate on both CT and ST is 0, with a mean 16 on CT, and a mean 15.5 on ST. The errors that occurred are outlined in Table 5.

Table 5. Experiment II: Assembly errors by treatment

| Treatment | Incorrect selection of variants | Omissions | Incorrect fixation/ | Picked up poor quality parts | Wrong counting | Errors Made | Failed Parts |
|---|---|---|---|---|---|---|---|
| ST | 19 | 4 | 5 | 2 | 1 | 31 | 29 |
| CT | 20 | 0 | 6 | 6 | 0 | 32 | 30 |

The data are discrete. The data and their logs are non-parametric. A Levene's test ($F$ = 0.2301, $p$ = 0.6317) indicates homogeneity of variances, therefore Wilcoxon Rank Sum is an appropriate statistical test. A Wilcoxon test showed that there is no significant difference ($S$ = 39902, $Z$ = –0.27989, $p$ = 0.7796) between the number of errors made by participants using either treatment. Given the large number of zeros in the dataset the non-zero sample size is small, the large-sample normal approximation might not be adequate, and it is therefore appropriate to compute the Exact Test. For the Exact Wilcoxon Test, the one-sided *p*-value is 0.4133, and the two-sided *p*-value is 0.8265. As the *p*-values are greater than or equal to the significance level of 0.05, the decision is to not reject the null hypothesis $H2_0$. ST does not prevent error from occurring.

There is however a significant difference in how the participants respond to error. In Table 6, it can be seen that of the 31 errors made in ST, 30 are detected and 29 corrected, leaving two errors. The first of the ST failed assemblies is as a result of wrong counting. This is detected by the system, but the participant ignored the system prompt and did not continue with the assembly and ended the experiment. The second failed ST assembly occurred due to distinguishing features of the part being obscured from the cameras. The machine learning algorithms incorrectly identify the part as being correct. In CT, 32 errors are made, and the participant detects 20 of them. Of that 20, 19 are corrected, with one having been detected but not corrected. Therefore, 13 of the 32 errors in CT remain in the products post-assembly process. These products fail the quality check.

Table 6. Experiment II: Response to assembly errors by treatment

| Treatment | Failed Parts | Errors Made | Errors Detected | Errors Corrected | Errors Remain |
|---|---|---|---|---|---|
| ST | 29 | **31** | 30 | 29 | 2 |
| CT | 30 | **32** | 20 | 19 | 13 |

For the two-sample Exact Wilcoxon test on Errors Corrected, the one-sided *p*-value is 0.0013, and the two-sided *p*-value is 0.0021. For the two-sample Exact Wilcoxon test on Errors Remaining, the one-sided *p*-value is <0.001, and the two-sided *p*-value is <0.001. As recommended in Divine *et al.* [12] further evidence such as the percentage showing improvement should be provided on trials which report Wilcoxon results, particularly in cases where the sample median is equal to zero. In this study, for the ST, 29 of 31 errors (93.5%) were corrected, whereas in CT only 19 of 32 (59.3%) were corrected. Using ST the final error is ($n_{ST}$= 2/200) is 1%, although not all of these errors had a negative impact on product quality, for example one was a counting error. In CT, the final error is ($n_{CT}$ = 13/200) is 6.5%. Participants using the ST are more likely to correct the error than those using CT, leaving fewer errors remaining in the assemblies. As the *p*-values of the Wilcoxon tests are less than or equal to the significance level of 0.05 the decision is to reject the null Hypothesis $H3_0$. The use of an AS4.0 did result in lower rates of failed quality products.

### A. Discussion

The key results of the experiments are now summarised with reference to the study objectives.

#### 1) Impact on detection of human errors

The first hypothesis *H1: The use of an AS4.0 will detect errors made in assembly* is investigated through Experiment I a laboratory simulation. There are several other studies that prove that AI can detect parts and assemblies. However, they are often in fixed locations on a work station table, and deal with larger sized parts easily identified. In this research an assembly that can occur anywhere in a three-dimensional space of the work station with small parts is replicated. The pen assembly uses very small parts, and introduces similar variants where the difference between the parts is measured in millimeters. The results prove that the AS4.0 can identify error on this specific use case. A wide range of error types are tested. Of the one hundred samples, of which 50 are correct, and 50 are incorrect, the data-fusion approach has an accuracy of 0.94, precision at 0.98, and recall of 0.92. These values can be further improved by introduction of new data through metrology. The original four False Negatives arose from not measuring the length of the spring or the ink chamber protruding out of the barrel. Measuring length when the part can be held nearer or further away from the camera, and/or held at an angle where the full barrel may not be measured is difficult. This can be corrected by use of a digital calipers to measure the length of the part. Once the digital calipers are added the barrel with the ink chamber and spring were measured, and the data integrated into the prediction. The False Negatives were removed. Using these data, accuracy is recalculated as 0.98, precision at 0.98, and recall is 1. However, the introduction of the measurement by the calipers changes the assembly process. This is an experiment only, but in a real assembly, the Process Engineer responsible would make the decision as to whether the False Negatives can be allowed, whether more data should be fed to the model for training, or whether to alter the process to introduce some metrology. Howsoever, it is concluded therefore that the data-fusion approach used in the AS4.0 can detect human error. The null hypothesis $H1_0$ is rejected.

#### 2) Impact on Performance: Reduction of Human Error

The second and third Hypotheses, H2*: The use of an AS4.0 will reduce the amount of human errors*; and H3*: The use of an AS4.0 will result in lower rates of failed quality products* are investigated through Experiment II. This experiment allows for the maximum potential for precise measures of behaviour of human participants.

With respect to H2*: The use of an AS4.0 will reduce the amount of human errors*, no statistical difference is found between the two treatments. The AS4.0 did not prevent human error from occurring any more so than the AS. The median and mode error rate on assembly is 0 for both treatments. Riedel *et al.* [6] found that in using AS4.0 *"...errors by omission...quantitative errors...and execution errors…are effectively prevented by the assembly assistance system. As the system allows the execution of the following work step only if the current work step is visually detectable, errors by omission are actively prevented*". However, in Experiment II herein, it is found that the participants can choose to ignore the assistance of the system. In fact, although the system indicated a "wrong counting" error, one participant did not respond to it. The results therefore differ from the conclusions of Riedel *et al.* [6], in that error can be identified, and flagged to the human, but not actively prevented. It is up to the human to use the information or not.

Of particular interest in the study of error is the impact on quality, as is evaluated in *H3: The use of an AS4.0 will result in lower rates of failed quality products*. The AS4.0 did not stop the human making error. However, the extent to which those errors cause failed quality products can be reduced. In the AS4.0 herein, the error is observed and flagged to the participant. There is a statistically significant difference in the error detected and corrected between ST and CT. Thus, whilst the number of errors made is not reduced, the number of failed quality parts is significantly reduced. AS4.0 had a positive impact on the number of high quality products being produced.

*3) Other findings: Impact on time*

The participants were observed, their behaviour noted and timings recorded. Using a t-Test it was found that the Cycle Times in ST (mean 8.65, median 7) are shorter than the CT group (mean 13.96, median 11) ($t(198) = -6.1217, p = 0.0001$) with a difference of -5.31. The use of ST positively impacted the speed of work. There is no statistical difference between the Execution Times between the treatments (Wilcoxon Rank Sum $Z = -0.78019$, $p = 0.4353$). The participants took the same amount of time to complete the assembly regardless of treatment with ST median of 4 s mean 5.75 s, and CT median 4 mean 6.33. An explanation for the increased Cycle Time in CT is consultation time. In CT, the participant had to read text instructions, and had to take time to press the "confirm assembly" button on the touchscreen. In ST, instructions are provided by video which may be quicker to process than text, and the "confirm" state is indicated by the machine learning algorithms not by the participant. The consultation time for CT is a median of 5 s and for ST a median 2 s, with Consultation Time ranges CT 0 s–54 s, and ST 1 s–25 s. ST had lower consultation times than CT. However, it is cautioned that participants only completed five assemblies each. It is to be expected that Consultation Times will decrease with repeated use. Of particular interest is the Consultation Time range, CT 0 s–54 s, with a low value of 0. In CT, four of the ten participants stopped consulting the system during the assembly, on a total of 71 of the 200 assemblies. CT requires that the participant indicate that the work is complete by selecting the confirm button on the touchscreen monitor. In these 71 assemblies the participant does not confirm the assembly, so the inherent quality control check is not completed. This issue did not occur in ST where 100% of all assemblies are checked by the machine learning algorithms and a "correct" or "wrong" flag recorded. Parmentier *et al.* [13] warn that humans often ignore assembly instructions which may lead to error. This is found to be true in these experiments. The human did ignore the assembly instructions, although in these experiments ignoring the instructions did not lead to higher rates of error than in assemblies where instructions are consulted. However, it is cautioned that the assembly use case herein is a simple one, and the lack of Consultation Time on more complex assemblies may have greater impact on error.

## IV. Conclusion

This research set out to validate an Assembly System 4.0 and its impact on reducing error and improving human performance. There are a number of research outcomes, which are evaluated in two different experiments. Three hypotheses are posed that express a possible relationship between the treatments and the research outcomes.

H1. The use of an AS4.0 will detect errors made in assembly.

H2. The use of an AS4.0 will reduce the amount of human errors made in assembly.

H3. The use of an AS4.0 will result in lower rates of failed quality products.

In the first experiment, a laboratory simulation, an effort is made to identify and impose control over assembly and assembly errors. One hundred tests were conducted generating 700 data points, half of which are with error, and half of which are without error. A wide range of error types are tested of the types Error by Omission, Error by Confusion, and Execution Errors. A data-fusion approach is used, where object detection, classification, and segmentation are conducted using YOLOv8. Classification robustness is tested and false predictions are eliminated using Inception v3. Multiple cameras view the object being assembled from different angles. The image feed from each camera are processed and the classification robustness are integrated to produce consistent and accurate information. This approach had an accuracy of 0.94, precision at 0.98, and recall of 0.92. These values are further improved by introduction of new data through metrology, leading to new calculations of accuracy 0.98, precision 0.98, and recall of 1. With respect of Hypothesis 1, the AS4.0 is found to have utility in correctly predicting error.

In the second experiment, two different treatments are used to determine on which research outcomes a treatment has an effect—a system treatment and a control treatment. Two assembly systems were built. The first treatment is the Assembly System 4.0, an assembly system that uses a fusion of machine learning techniques to augment the operator in assembly. The second treatment is a traditional Assembly System design, which provides a web-page based interface to the assembly instructions and is presented to the participant via a touchscreen. Two independent samples are created. Twenty participants took part of which ten were assigned to the system treatment, and ten to the control treatment. The participants assembled five products, each of which had four assembly steps. In total, four hundred observations are recorded. Data is generated by the system and by observation, creating 9,200 data points.

With respect to Hypothesis 2, neither treatment is shown to have a positive impact on reducing the number of human errors made. The participant continued to make error. However, with respect to Hypothesis 3, the use of an Assembly System 4.0 had a statistically significant impact on the number of quality products produced. A quality product is produced, when either the assembly is executed correctly first time, or if an error has been made, and it is detected and reworked by the human. Using the Assembly System 4.0, ST, 93.5% of errors were corrected, whereas using the traditional Assembly System, CT, 59.3% were corrected. Using ST the final error rate is 1%, although not all of these errors had a negative impact on product quality, and in CT, the error final rate is 6.5%. Participants using the ST are more likely to correct the error than those using CT, leaving fewer errors remaining in the assemblies.

It is also noted that the Assembly System 4.0 had a positive impact on performance in terms of cycle time and consultation time. As consultation with the system, and the performance of in-line quality control checks conducted at each stage of the assembly process have an impact on error, these actions missing from a traditional Assembly System are potentially important for error outcome.

The results of these experiments are statistically significant with an alpha <0.05, and provide evidence to assist practitioners in understanding the value-add of such a system in reducing error.

### A. Limitations

Sample bias may have occurred in Experiment II. The participants are from a student body with an interest in manufacturing solutions. They may not reflect the manufacturing operator population. The students had no prior experience of assembly on a busy factory floor, and how instructions are typically retrieved and interpreted on a factory floor.

### B. Future Work

Future work should include evaluation in context. Thus, how the system reduces the environmental pressures inherent in the workplace for the operator can be evaluated. This will further substantiate Assembly Systems 4.0 by providing evidence that they have utility. As users are key to successful adoption of the system research should be conducted with users so as to understand their perspective of the system.

## Conflict of Interest

JH is named as an inventor on a patent application for an Assembly System 4.0. The authors declare no conflict of interest.

## Author Contributions

JH conducted the research, analyzed the data, and wrote the paper. PV and AR contributed to the final version of the manuscript. All authors approved the paper.

## Funding

This research was funded by Enterprise Ireland and European Regional Development Fund (ERDF) under Ireland's European Structural and Investment Funds (ESI) Programmes 2014–2020, grant number IP20170566.

## Acknowledgment

The authors wish to thank Jeremiah Scully, Iram Ashad and David Byrne.

## References

[1] J. Wolfartsberger, J. Haslwanter, and R. Lindorfer, "Perspectives on assistive systems for manual assembly tasks in industry," *Technologies*, vol. 7, no. 1, 12, Jan. 2019. doi: 10.3390/technologies7010012

[2] I. Badr and B. Vogel-Heuser, "Taxonomy of faults influencing the quality of automation production systems," *IFAC-PapersOnLine*, vol. 51, no. 10, pp. 95–102, 2018. doi: 10.1016/j.ifacol.2018.06.243

[3] A. Dolgui, F. Sgarbossa, and M. Simonetto, "Design and management of assembly Systems 4.0: Systematic literature review and research agenda," *Int. J. Prod. Res.*, vol. 60, no. 1, pp. 184–210, Jan. 2022. doi: 10.1080/00207543.2021.1990433

[4] A. Miqueo, M. Torralba, and J. A. Yagüe-Fabra, "Lean manual Assembly 4.0: A systematic review," *Appl. Sci.*, vol. 10, no. 23, 8555, Nov. 2020. doi: 10.3390/app10238555

[5] C. Stockinger, L. Polanski-Schräder, and I. Subtil, "The effect of information level of digital worker guidance systems on assembly performance, user experience and strain," *Appl. Ergon.*, vol. 106, 103896, Jan. 2023. doi: 10.1016/j.apergo.2022.103896

[6] A. Riedel *et al.*, "A deep learning-based worker assistance system for error prevention: Case study in a real-world manual assembly," *Adv. Prod. Eng. Manag.*, vol. 16, no. 4, pp. 393–404, Dec. 2021. doi: 10.14743/apem2021.4.408

[7] J. Böllhoff, J. Metternich, N. Frick, and M. Kruczek, "Evaluation of the human error probability in cellular manufacturing," *Procedia CIRP*, vol. 55, pp. 218–223, 2016. doi: 10.1016/j.procir.2016.07.080

[8] T. Nakajo and H. Kume, "The principles of foolproofing and their application in manufacturing," Reports of Statistical Application Research, 1985.

[9] J. P. Sondermann, *Poka Yoke, Prinzip und Techniken für Null-Fehler_Prozesse*, Munich: Hanser, 2013.

[10] G. E. Noether, "Sample size determination for some common nonparametric tests," *J. Am. Stat. Assoc.*, vol. 82, no. 398, pp. 645–647, Jun. 1987. doi: 10.1080/01621459.1987.10478478

[11] T. Keller, C. Bayer, P. Bausch, and J. Metternich, "Benefit evaluation of digital assistance systems for assembly workstations," *Procedia CIRP*, vol. 81, pp. 441–446, 2019. doi: 10.1016/j.procir.2019.03.076

[12] G. Divine, H. J. Norton, R. Hunt, and J. Dienemann, "A review of analysis and sample size calculation considerations for wilcoxon tests," *Anesth. Analg.*, vol. 117, no. 3, pp. 699–710, Sep. 2013. doi: 10.1213/ANE.0b013e31827f53d7

[13] D. D. Parmentier, B. B. V. Acker, J. Detand, and J. Saldien, "Design for assembly meaning: A framework for designers to design products that support operator cognition during the assembly process," *Cogn. Technol. Work*, vol. 22, no. 3, pp. 615–632, Aug. 2020. doi: 10.1007/s10111-019-00588-x