

RESEARCH ARTICLE OPEN ACCESS

World of ScoreCraft: Novel Multi-Scorer Experiment on the Impact of a Decision Support System in Sleep Staging

Benedikt Holm¹  | Arnar Óskarsson¹ | Björn Elvar Þorleifsson¹ | Hörður Þór Hafsteinsson¹ | Sigríður Sigurðardóttir²  |
 Heiður Grétarsdóttir² | Kenan Hoelke² | Gabriel Marc Marie Jouan¹ | Thomas Penzel³  | Erna Sif Arnardóttir^{2,4}  |
 María Óskarsdóttir^{1,5} 

¹School of Technology, Department of Computer Science, Reykjavik University, Reykjavík, Iceland | ²School of Technology, Sleep Institute, Reykjavik University, Reykjavík, Iceland | ³Sleep Medicine Center, Charité—Universitätsmedizin Berlin, Berlin, Germany | ⁴The National University Hospital of Iceland, Landspítali, Reykjavík, Iceland | ⁵School of Mathematical Sciences, University of Southampton, Southampton, UK

Correspondence: Benedikt Holm (benedikthth@ru.is)

Received: 8 January 2025 | **Revised:** 18 May 2025 | **Accepted:** 27 May 2025

Funding: This research was supported by the European Union's Horizon 2020 research and innovation programme (grant agreement 965417). Benedikt Holm received funding from the ErasmusPlus programme to travel to Penzel's lab in Berlin.

Keywords: artificial intelligence | decision support system | scoring accuracy | sleep staging

ABSTRACT

Manual scoring of polysomnography (PSG) is a time-intensive task, prone to inter-scorer variability that can impact diagnostic reliability. This study investigates the integration of decision support systems (DSS) into PSG scoring workflows, focusing on their effects on accuracy, scoring time and potential biases toward recommendations from artificial intelligence (AI) compared to human-generated recommendations. Using a novel online scoring platform, we conducted a repeated-measures study with sleep technologists, who scored traditional and self-applied PSGs. Participants were occasionally presented with recommendations labelled as either human- or AI-generated. As the goal of this study was to isolate the effect of perceived recommendation sources on scorer behaviour, all recommendations were human-generated. We found that traditional PSGs tended to be scored slightly more accurately than self-applied PSGs, but this difference was not statistically significant. Correct recommendations significantly improved scoring accuracy for both PSG types, while incorrect recommendations reduced accuracy. No significant bias was observed toward or against AI-generated recommendations compared to human-generated recommendations. These findings highlight the potential of DSSs to enhance PSG scoring reliability. However, ensuring the accuracy of the suggestions is critical to maximising its benefits. Future research should explore the long-term impacts of DSS on scoring workflows and strategies for integrating AI in clinical practice.

1 | Introduction

As sleep disorders and sleep issues are extremely prevalent in society (Arnardóttir et al. 2016; Adams et al. 2017; Zeng et al. 2020), sleep technologists are more important today than ever. The sleep technologist is responsible for reviewing sleep studies, or PSG, and manually annotating (also known as scoring) sleep stages and events such as movement, arousal and

apnea (American Academy of Sleep Medicine 2023). Medical doctors subsequently use this analysis for diagnosis.

A PSG is an overnight collection of biometric signals, such as various respiratory signals, electrooculogram (EOG), electroencephalogram (EEG) and electromyogram (EMG), to name a few (American Academy of Sleep Medicine 2023). The scoring of a PSG is a time-consuming task, which takes up to 2 h to score a

Abbreviations: AI, artificial intelligence; DSS, decision support system.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Journal of Sleep Research* published by John Wiley & Sons Ltd on behalf of European Sleep Research Society.

single 8-h PSG (Rayan et al. 2023) and can significantly strain the sleep technologist and induce scoring fatigue. Alongside being laborious and time-consuming, the scoring of a PSG can also vary significantly between sleep technologists, with disagreement on sleep staging being as high as 14% (Nikkonen et al. 2024) and respiratory events as high as 34.6% (Rosenberg and Van Hout 2014), and obstructive apnea severity classifications differing in 66% of cases depending on the scorer (Pitkänen et al. 2024).

To reduce the need for a sleep technologist to set up and monitor a traditional PSG and to eliminate the requirement to sleep in a sleep laboratory overnight, new PSG equipment has been developed for home use. This multichannel frontal PSG enables patients to apply the device themselves and sleep comfortably in their own beds. These self-applied PSG devices are currently undergoing validation, and a recent device demonstrated a success rate of approximately 90% in enabling effective self-application and accurate data collection (Ferretti et al. 2022).

The rise of artificial intelligence (AI) has not left the scoring process behind, with various automatic algorithms being designed to automate and speed up the work of sleep technologists by detecting sleep stages (Wara et al. 2024) and apnea (Moridian et al. 2022; Mostafa et al. 2019) to name a few tasks. These algorithms have great potential to assist in the scoring process; however, they cannot be treated as a drop-in replacement for the human expert since they are incapable of adapting and adjusting to the evolving standards of care as the human experts are. They also need to be rigorously trained for different issues that may arise in diverse application contexts and be aware of possible biases that occur in AI based on sex, age, comorbidities and other factors (Varsha 2023).

To meet the need for systems that aim to utilise AI as a tool for the human expert instead of as a replacement, decision support systems (DSS) have been designed to provide interactive tool sets to assist experts in making decisions and solving unstructured or semi-structured tasks (Sprague 1980). Garg et al. (2005) found that in 64% of cases, DSS or similar systems considerably impacted clinician performance. Articles on DSS have been widely written in the field of sleep research; however, the literature covers mostly automation of individual tasks but does not examine the effect of the inclusion of DSS into the workflows of sleep technologists in terms of accuracy or time taken to score sleep recordings. Furthermore, the effects of integrating DSS and AI into sleep technologists' workflows can provide significant advantages in terms of speed and accuracy, but to be integrated effectively requires building trust toward the AI (Asan et al. 2020).

Along with the effects as mentioned earlier, an important aspect to measure is the potential impact of the DSS in changing the behaviour of the human expert or how the professionals whose toolset is augmented with AI might, for example, become complacent and default to the AI recommendations (Parasuraman and Manzey 2010). Complacency toward AI can be defined as a tendency of the user to not appropriately scrutinise the results of the automated tools. The tendency toward complacency is complicated to analyse but has been shown to be linked to the transparency of the AI system, as well as how well the expert expects

the AI to perform (Harbarth et al. 2024). Another important aspect of integrating AI is the concept of 'clinical acquiescence', defined by Holm et al. (2024), which refers to the willingness to adopt AI assistance in clinical workflows.

There is a noticeable gap in the literature on the effects of integrating DSS into the workflows of sleep technologists. Most of the existing literature focuses on the accuracy of the algorithms designed to automate sleep-scoring tasks (Rusanen et al. 2024; Holm et al. 2024). However, the impact assessment of such algorithms on expert performance is a key component that is usually missing. We propose to leverage this by studying the changes introduced by human experts whose toolsets have been augmented with DSS. In more detail, we aim to investigate the effects of integrating AI into the work environment.

This work investigates the effects of introducing recommendations in scoring sleep stages. We further measure the effects of the recommendation presentation and correctness on the accuracy and speed of the human expert.

This study used a repeated measures design with two conditions to collect a consensus scoring for 1 h of traditional and self-applied PSG. The main conditions being researched were the effect of recommendations presence and study type, as the objective of this study is to research the effects of recommendations on the sleep staging process. The study also examines the effect of the type of scoring recommendations (human or AI) on the scoring process, counterbalancing the recommendations by only showing recommendations for one session of each PSG type. Hence, three factors were being studied: type of sleep study, presence of recommendations and type of recommendations. Their significance is measured in terms of scoring accuracy or correctness and time.

2 | Methodology

For this study, the scoring sessions were limited to a single hour (120 epochs), chosen from a data set that was recorded simultaneously using traditional PSG and self-applied PSG equipment. The hour was chosen for good-quality signals, and the hypnogram featured multiple transitions between sleep stages. The sessions were limited to 1 h to focus on collecting a greater diversity of scorings and to prevent scoring fatigue from affecting sleep technologists during the process.

2.1 | Platform

The scoring collection was performed with the MicroNyx online scoring platform (Holm 2023), which allows secure online scoring of PSGs. The MicroNyx platform enables measuring difficult-to-obtain features, such as the decision-making time, change of mind and more aspects of the scoring process. In preparation for this study, multiple sleep technologists were recruited before the experiment to validate and provide recommendations and feedback on the scoring interface and signal filtering to measure the impact of the new MicroNyx system on the scoring in a Co-Design process. This was repeated several times over a few months until each sleep

technologist at Reykjavik University could reliably score with 80% consensus with a pre-existing scoring, or roughly around the 86% expected agreement of sleep technologists (Nikkonen et al. 2024). MicroNyx allows for the creation and the scoring of so-called 'scoring sessions', variable-length signal segments containing signals that sleep technologists can then score for research purposes. The signals, filtering and data source can be customised and tailored to different research purposes.

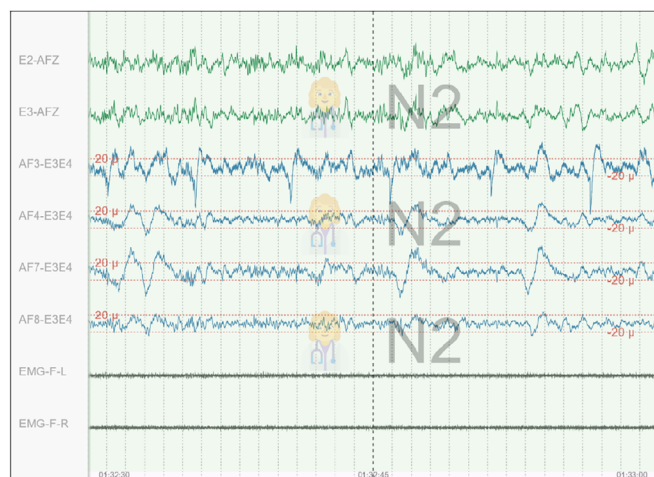
2.2 | Recommendations

In 50% of epochs, participants were presented with a recommendation for a sleep stage. The recommendation rate of 50% was chosen to provide an equal amount of epochs with and without scoring recommendations. To investigate the presence of any potential bias against automatic algorithms, the recommendations were sourced from a human scoring for the PSG and self-applied PSG studies, respectively. For consistency, the recommendations were sourced from a single sleep technologist who scored both the traditional and the self-applied PSG. This was done in order to eliminate bias due to the scorer. Despite all scoring being sourced from the same human sleep technologist, the recommendations were either presented as being from a human sleep technologist or an automatic AI scoring system. This is henceforth referred to as the recommendation presentation. The ratio of human versus AI recommendations was equal to ensure equal representation.

Both accurate and deliberately incorrect recommendations were implemented to evaluate the influence of recommendations on scoring behaviour. The rate of incorrect recommendations was



FIGURE 1 | Map of sleep stages to deliberate misclassification for scoring recommendations.



(a) Recommendation presented to be from a human

set to 20%, aligning with the inter-scorer variability reported by Nikkonen et al. (2024) and emulating the error rate of sleep technologists. To ensure realistic incorrect recommendations, a sleep-stage map was developed using results from a multicentric consensus scoring study (Nikkonen et al. 2024), which identified the most common misclassifications by sleep technologists. Figure 1 illustrates the most common misclassifications, providing insight into how incorrect recommendations were designed to reflect typical scoring disagreements.

Recommendations presented as being from humans were represented with a scientist emoji (see Figure 2a). In contrast, the recommendations we presented as being from an AI had the robot emoji, as shown in Figure 2b. The recommendations were intentionally made prominent and easily noticeable to sleep technologists, without obstructing or distorting the underlying signal. This design ensured that participants could clearly identify whether a recommendation was labelled as human- or AI-generated, enabling us to assess whether their behaviour reflected any detectable source-based bias. The post-study survey included questions about the visibility of the recommendations to ensure that the participating sleep technologists could easily spot the recommendations and tell the difference between human and AI recommendations.

2.3 | Data Setup

The data for this study was selected from a 'double-setup' dataset, where a traditional PSG recording along with a self-applied PSG setup was placed on the participants, and the two types of PSG were recorded simultaneously (Rusanen et al. 2024).

The sleep technologists scored both types of PSG using the MicroNyx web scoring interface, which supports flexible and customizable signal selection and filtering. Scoring guidelines for required signals and their appropriate filtering were followed to ensure consistency with established methodologies and software for both traditional and self-applied PSG. This subsection is divided into two parts, detailing the signals and filtering options for traditional PSG and self-applied PSG, respectively.



(b) Recommendation presented to be from an AI

FIGURE 2 | Comparison of recommendations presented as human versus AI on different epochs.

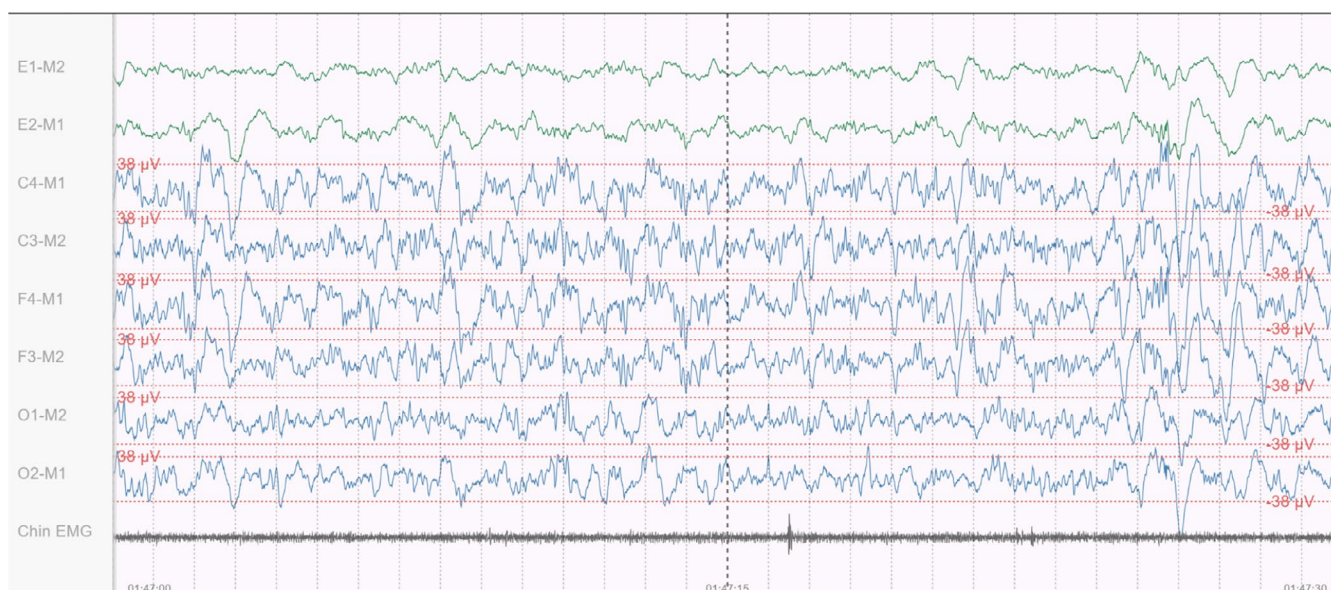


FIGURE 3 | The MicroNyx scoring interface displaying a traditional PSG.

2.3.1 | Traditional PSG

The traditional PSG setup presented to the sleep technologists was based directly on the AASM recommendations for scoring sleep stages, utilising the appropriate signals and filters as listed in the guidelines (American Academy of Sleep Medicine 2023). An illustration of the PSG signals is provided below.

The signals the sleep technologists used to score the traditional PSG were the EEG signals C4-M1, C3-M2, F4-M1, F3-M2, O1-M2, O2-M1, the EOG signals, E1-M2, E2-M1 and the chin EMG.

The EEG signals were filtered using a 0.5–35 Hz bandpass filter. Each EOG signal was processed through a 0.3–35 Hz bandpass filter and sampled at 200 Hz. The chin EMG signal was passed through a 10 Hz high-pass filter and sampled at 200 Hz.

Figure 3 shows how the signals from the traditional PSG were presented to the sleep technologists in the MicroNyx scoring interface.

2.3.2 | Self-Applied PSG

The self-applied PSG signal used by the sleep technologists to score the self-applied PSG were the EEG signals AF3-E3E4, AF4-E3E4, AF7-E3E4, AF8-E3E4 and the EOG signals E2-AFz, E3-AFz. The EEG signals were filtered using a 0.5–35 Hz bandpass filter and sampled at 200 Hz. The EOG signal was processed through a 0.3–35 Hz bandpass filter and sampled at 200 Hz.

Since the self-applied PSG setup did not include a traditional chin EMG, the E1 signal referenced against the E3 signal, along with the E2 signal referenced against the E4 signal, were used as stand-ins for the EMG signal, with a 10 Hz high-pass filter applied to produce left and right EMG signals, respectively.

Figure 4 shows how the signals from the self-applied PSG were presented to the sleep technologists in the MicroNyx scoring interface.

2.4 | Reference Standard

After the scoring collection, a so-called reference standard was created, a majority-vote scoring for each epoch, which could be used to compare scorings. To have a stable, more reliable hypnogram to compare scorings to than the single-scorer hypnogram, the user scorings for the traditional PSG sessions that did not have recommendations present were used to create a consensus scoring referred to henceforth in this work as the reference standard. This was done to alleviate two major issues. The first issue is that the single-scorer hypnogram was created on different software and is thus not perfectly comparable to the scorings generated on the platform used for this study. The second issue that the reference standard solves is that the scoring of a self-applied PSG is not as validated as a traditional PSG. By creating a consensus scoring for the traditional PSG that can be compared against the scorings for the self-applied PSG, the difference in the scoring of traditional and self-applied PSG is more pronounced. This work uses the reference standard to compare the participant scorings and the recommendations. We refer to sleep technologists' scorings as *accurate* and recommendations as *correct* if they respectively match the reference standard. The words are not considered interchangeable in this work since recommendations cannot be considered 'accurate', as they are not themselves scorings.

2.5 | Recruitment of Participants

The study was heavily dependent on the participation of sleep technologists. To recruit sleep technologists as participants in the scoring collection, an email invitation was sent to 49 sleep technologists affiliated with the Sleep Revolution (Arnardottir

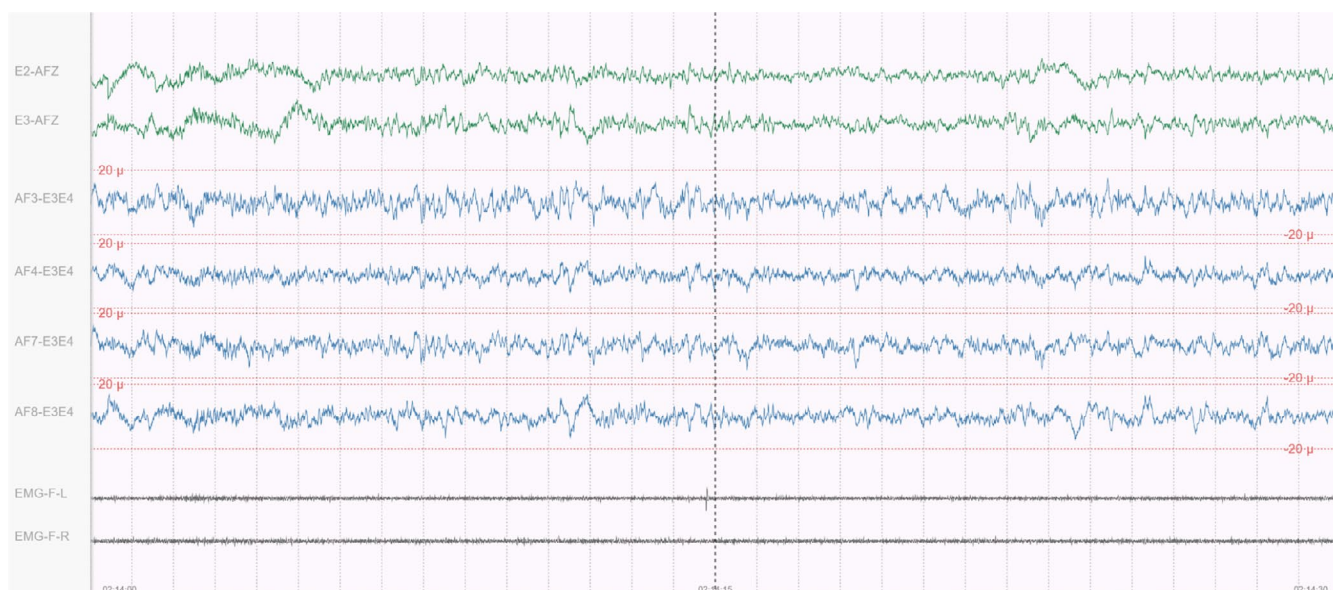


FIGURE 4 | Self-applied polysomnography in the MicroNyx scoring interface.

et al. 2022) project. In addition, the study was advertised at conferences in the European Respiratory Society (ERS) and the European Sleep Research Society (ESRS) in September 2024. An invitation was also emailed to European Society of Sleep Technologists members via a newsletter. In total, 16 sleep technologists across Europe participated in the study.

2.6 | Study Procedure

Each technologist was instructed to complete a 10-min (20-epoch) tutorial session in which they were shown how to navigate the scoring interface, how recommendations appeared, and how to score using the MicroNyx interface successfully. After the tutorial, each technologist was directed to complete two scoring sessions, one for PSG and one for self-applied PSG, where they scored exactly 1 h of sleep data per session. After those two sessions, the sleep technologists were instructed to wait a week before scoring two additional sessions for the same recordings. The waiting time was instructed to prevent familiarity with the recordings.

The same 1-h segment was used for both the traditional and self-applied scoring sessions, as the two PSG types were recorded in parallel on the same individual. This ensured that participants scored simultaneous data captured by different equipment, allowing for direct comparison between traditional and self-applied PSG without variability in underlying sleep physiology.

After successfully completing the four sessions, a link to a short post-study survey was presented to the sleep technologists. The post-study survey aimed to gauge how each sleep technologist perceived the MicroNyx platform for its ease of use and ability to score sleep stages on the scoring interface. The post-study survey questions are provided in Appendix A.

The MicroNyx platform ensured that each participant completed one traditional PSG session and one self-applied PSG

session in randomised order, followed by a one-week break, after which they repeated the process with a second pair of sessions. Recommendations were presented in only one of the two traditional sessions and one of the two self-applied sessions, ensuring that if recommendations were provided in the first traditional session, they would not be shown in the second traditional session, and the same rule applied to the self-applied sessions. Recommendations are covered in more detail in the following section.

2.7 | Analysis

Data analysis was performed using Python and R, with tools chosen to suit the specific requirements of each test. In Python, the Pandas library (McKinney 2010) was used for data preparation and organisation. For simpler single-variable analyses, the SciPy library (Virtanen et al. 2020) was employed to conduct hypothesis testing through paired *t*-tests, assessing significant differences in decision-making time and accuracies under different conditions, assuming normality. When the normality assumption was unmet, the Mann–Whitney test (Mann and Whitney 1947) was applied as a non-parametric alternative. A significance level of $\alpha = 0.05$ was used throughout. To investigate the relationship between scoring variables and decision-making time, the R programming language (R Core Team 2023) and the ARTool library (Kay et al. 2025) were utilised to perform an Aligned-Rank-Transform Analysis of Variance (ART ANOVA). ART ANOVA was chosen for its ability to accommodate the continuous nature of the decision-making time alongside categorical predictors such as recommendation correctness, presentation style and PSG type, which do not satisfy the assumptions of standard ANOVA. A generalised linear model was applied to analyse scoring accuracy, accounting for the binomial distribution of the dependent variable. This approach enabled the evaluation of categorical predictors' main effects and interactions, such as recommendation correctness, presentation style and PSG type, while respecting the constraints of binary data.

3 | Results

This section presents the results of this study, including the aggregate performance of sleep technologists when scoring traditional and self-applied PSG and the granular effects of recommendations on scoring accuracy and time. The analysis highlights differences between user-level and epoch-level outcomes, focusing on the impact of recommendation correctness and presentation. Cumulatively, the 16 participating sleep technologists completed 64 scoring sessions, producing 9158 individual scorings for the total 240 epochs in the scoring sessions. To create a reference standard, a majority vote approach was taken using the scorings of the 16 participants. This was used to assign a unique label to each epoch. To generate the reference standard, we applied a majority-vote procedure across multiple scorings. The original hypnogram used as the source for generating recommendations aligned with this reference standard with an accuracy of 76.23%. However, to simulate imperfect decision support, we intentionally introduced a 20% error rate into the recommendation set. As a result, the overall agreement between the recommendations and the reference standard decreased to 67.57% for the traditional PSG and 57.64% for the self-applied PSG. This indicates that, after introducing controlled errors, approximately 32% and 42% of the recommendations were ultimately incorrect.

When analysing the decision-making time, the time per epoch was obtained by calculating the time difference of the creation timestamps of successive scorings. A log transformation was applied to the time-per-epoch data to remove statistical outliers, and an interquartile range filtering with a threshold of 1.5 was used to remove unrealistically long decision-making times, potentially stemming from sleep experts standing up or getting distracted from the scoring process. The filtering step identified 598 outliers, with a mean time-per-epoch of 3210.6s and a standard deviation of 77646.7s, indicating that the filtered values deviated significantly from the remaining decision-making times. The remaining data used for the analysis had a mean scoring-per-epoch time of 2.0s and a standard deviation of 1.9s.

3.1 | Participants

In total, 16 sleep technologists participated in the study, successfully completing four scoring sessions. Of the 16 sleep technologists, 13 answered the post-participation questionnaire. The questionnaire (see Appendix A) showed that most sleep technologists felt confident in their ability to interpret and score signals, with over 84% selecting 6 or 7 on the confidence scale. Two participants (7.7% each) selected 2 or 5, indicating some variation in confidence levels. Of those who answered the questionnaire, all reported that they could easily see the recommendations and that it was easy to see if they were from a human or AI. The sleep technologists were from 11 different countries: Australia, Belgium, Finland, France, Germany, Guatemala, Iceland, Ireland, Italy, Portugal and Spain.

Of the various recruitment methods, the invitation sent to the relevant members of the Sleep Revolution (Arnardottir et al. 2022) yielded six sleep technologists, and the letter sent to the ESST yielded five. Two sleep technologists participated after

hearing about the study from colleagues or during a talk where the study was advertised.

3.2 | Aggregate Analysis of Scoring Accuracy and Time

To establish a baseline standard for scoring accuracy and time without the influence of recommendations, we filtered the data to exclude sessions that included recommendations from the analysis. When comparing the scoring accuracy of sleep technologists, 11 out of 16 technologists performed better in scoring the traditional PSG sessions than in scoring the self-applied PSG sessions (see in Figure 5 the accuracy achieved by each sleep technologist in both PSGs). Conversely, five sleep technologists scored the self-applied PSG sessions more accurately than the traditional PSG sessions. Overall, sleep technologists demonstrated a tendency for a slightly higher accuracy when scoring the traditional PSG sessions, achieving an accuracy rate of 85.7% compared to 81.0% for the self-applied PSG. The difference in accuracy was, however, not statistically significant (paired t -test: $p=0.098$, rank-sum: $p=0.15$).

The decision-making time was analysed similarly to the scoring accuracy. For the traditional PSG, the average decision-making time was found to be 2.0s, and for the self-applied PSG, 2.0s and is displayed in Figure 6, not statistically significant (paired t -test: $p=0.58$, rank-sum: $p=0.74$).

3.3 | Epoch-Level Effects of Recommendations on Accuracy

The sessions for both traditional and self-applied PSG were separated based on whether or not they included recommendations, and the scorings were then compared in terms of accuracy with the reference standard. The effect of correct recommendations on the overall accuracy of scorings for all scoring sessions can be seen in the heat map in Figure 7. Notably, the difference between accuracies for the self-applied PSG and the traditional PSG was more dramatic, with the baseline accuracy for self-applied PSG being 81.47%, but decreasing by

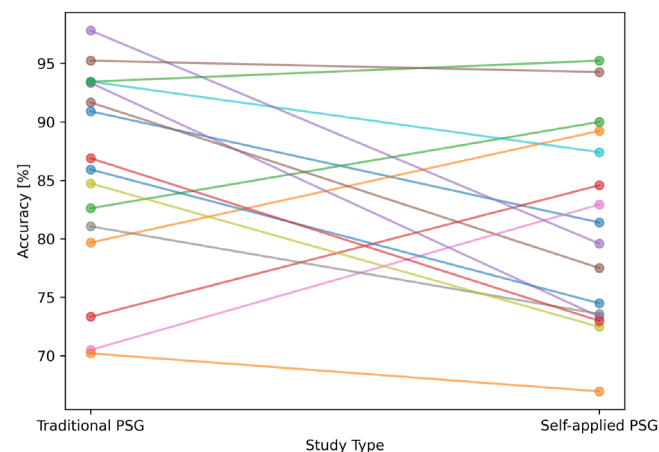


FIGURE 5 | Sleep technologist change in accuracy between traditional and self-applied PSG. Each line represents one sleep technologist.

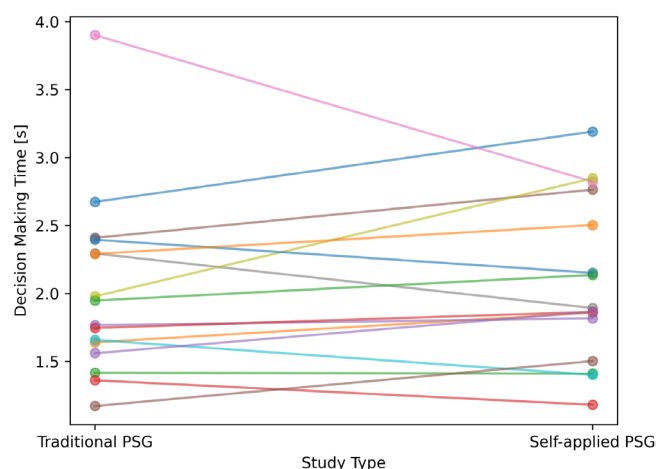


FIGURE 6 | Sleep technologist change in decision-making time between traditional and self-applied PSG. Each line represents one sleep technologist.

approximately 3.2% to 78.26% when recommendations were present. The accuracy difference was not statistically significant for the traditional PSG (t -test $p = 0.73$, rank-sum $p = 0.73$). However, for the self-applied PSG, the difference was found to be statistically significant (t -test $p = 0.006$, rank-sum $p = 0.006$).

When studied further by separating scorings based on their recommendation presence and correctness in Figure 8, the effect of correctness on recommendation becomes clearer. The baseline overall scoring accuracy on an epoch-by-epoch basis without recommendations remained at 85.38% for the traditional PSG sessions and 81.47% for the self-applied PSG sessions.

When faced with incorrect recommendations, sleep technologists assessed the traditional PSG with an accuracy of 82.13%, which is 3.26 percentage points lower than the baseline accuracy of 85.39% achieved when scoring the traditional PSG without any recommendations. This effect is even more pronounced

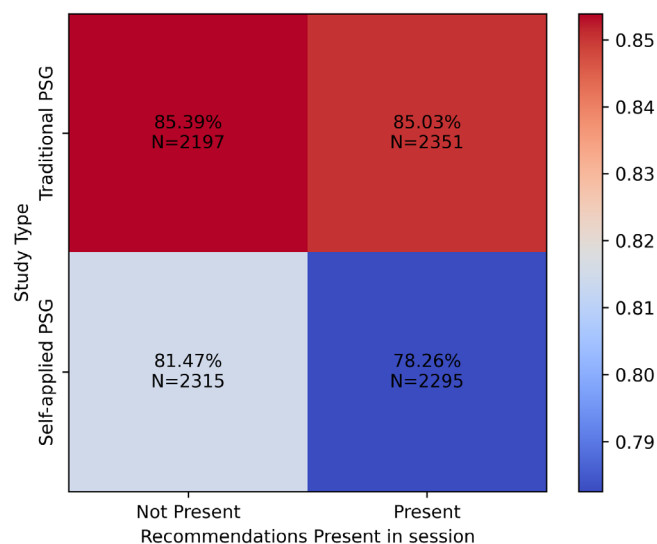


FIGURE 7 | Effect of recommendation presence on scoring session accuracy for traditional vs. self-applied PSG.

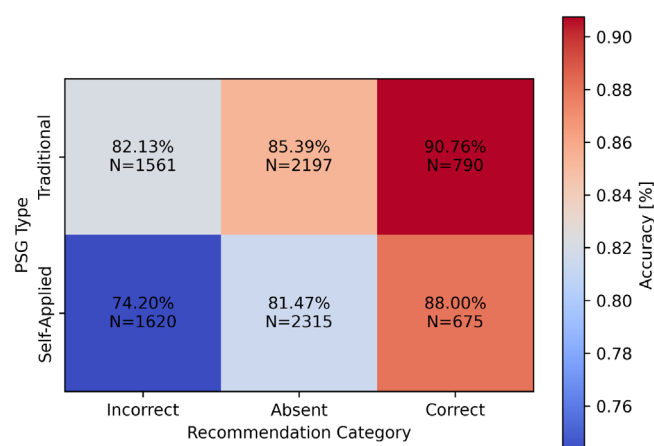


FIGURE 8 | Effect of recommendation correctness on scoring accuracy.

for the self-applied PSG, with the accuracy rate decreasing from the baseline 81.47% achieved when scoring the self-applied PSG without recommendations to 74.20% when faced with incorrect recommendations, making for a 7.27% decrease in accuracy. Correct recommendations had the opposite effect on scoring accuracy, with sleep technologists achieving 90.76% accuracy when scoring traditional PSG epochs featuring an accurate recommendation, resulting in a 5.37% increase in accuracy from the baseline scoring accuracy without recommendations. This positive effect was also observed for the self-applied PSG, where sleep technologists achieved 88% accuracy when presented with correct recommendations, up 6.53% from the 81.47% baseline scoring accuracy for self-applied PSG without recommendations.

The generalised linear model results can be seen in Table 1, which displays odds ratio (OR) change for each change in variables from the baseline (intercept) where the study type is a traditional PSG, the presentation is human, and the recommendation is correct. The baseline OR is 8.25, meaning that when study type, presentation and correctness equals the baseline, the accuracy of sleep technologists is 89.18%. The presentation and study type did not affect the scoring accuracy with statistical significance. However, the recommendation accuracy, on its own, had a significant effect on the scoring accuracy, lowering the accuracy of sleep technologists from the 89.18% baseline to 82.54%. The interaction between presentation and study type did not significantly affect the scoring accuracy, with AI recommendations when scoring self-applied PSG lowering the accuracy from the baseline by 10.32% to 78.85%. The final interaction that statistically significantly affected the scoring accuracy was for self-applied PSG when recommendations were incorrect, which lowered the scoring accuracy to 72.34%, or by 16.83%.

When plotted in a three-way line plot (see Figure 9), the interactions from Table 1 become more clear. For both traditional and self-applied PSG, incorrect recommendations decreased the scoring accuracy in line with the results from Figure 8. However, the negative impact of incorrect recommendations was not as dramatic for the traditional PSG as it was for the self-applied PSG. The presentation of correct recommendations had a paradoxical effect on accuracy with respect to the study types. For traditional PSG, human recommendations produced a mean accuracy of 90.43%, and

TABLE 1 | Generalised linear model linear regression results. The three-factor interaction term was included at first but was not significant. Thus, it was removed from the model.

	OR	2.5%	97.5%	Significance
Intercept	10.492	7.543	14.595	*
C(Presentation)[T.AI]	1.215	0.776	1.901	
C(StudyType)[T.Self-applied]	0.856	0.555	1.321	
C(Correctness)[T.False]	0.379	0.244	0.588	*
C(Presentation)[T.AI]:C(Study Type)[T.Self-applied]	0.661	0.404	1.082	
C(Presentation)[T.AI]:C(Correctness)[T.False]	1.062	0.654	1.723	
C(StudyType)[T.Self-applied]:C(Correctness)[T.False]	0.561	0.342	0.919	*

Note: Statistical significance is marked with an asterisk in the final column.

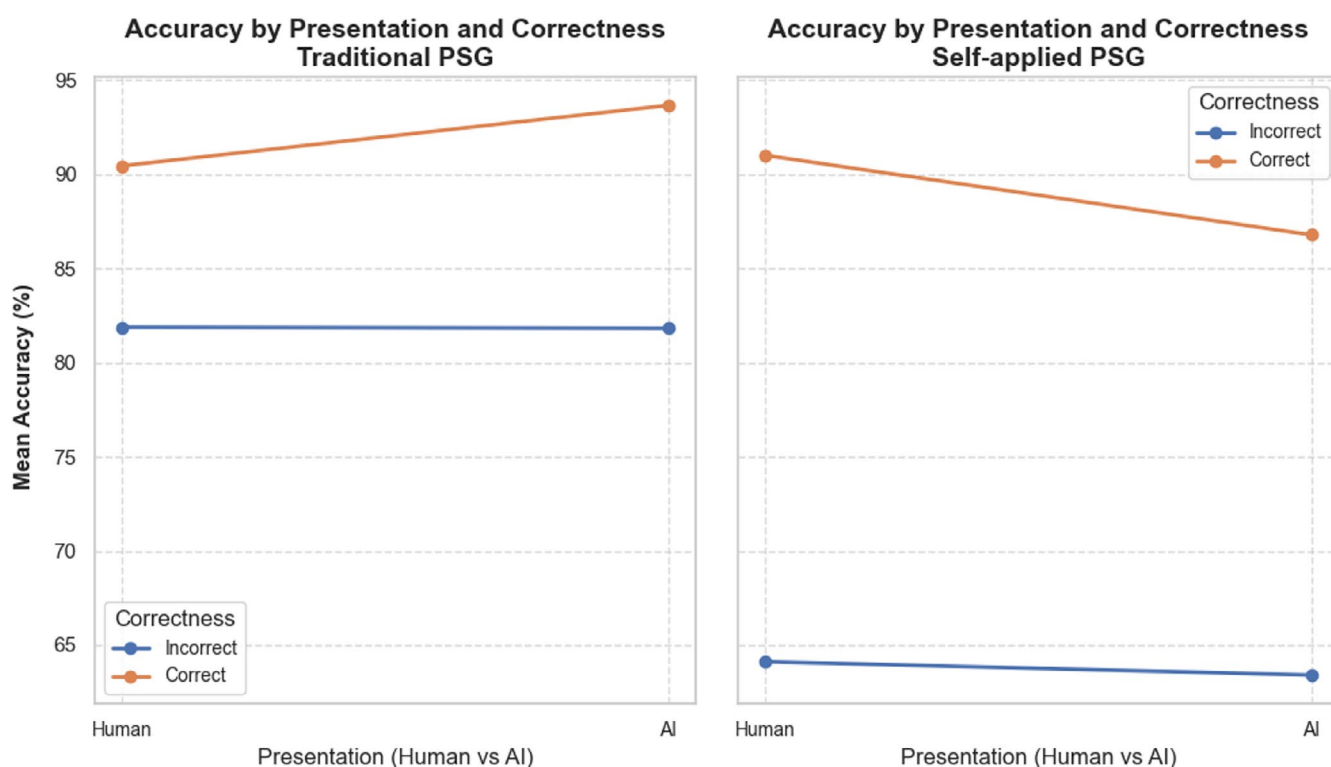


FIGURE 9 | Three-way line plot with grouped comparisons between the effects of study type, recommendation presentation and recommendation correctness on scoring accuracy.

AI recommendations produced a mean accuracy of 93.64%. Meanwhile, for self-applied PSG, human recommendations produced an average accuracy of 90.99% and AI recommendations produced a mean accuracy of 86.77%.

3.4 | Epoch-Level Effects of Recommendations on Decision-Making Time

The effects of recommendations on decision-making time were analysed using a method similar to the effects on accuracy. When separated based on study type and recommendation presence (see Figure 10), sleep technologists spent 1.9s scoring per epoch on average without recommendations, which rose to 2.0s per epoch when scoring with recommendations. This effect was

on the boundary of statistical significance (t -test $p=0.041$, rank sum $p=0.077$).

The average time per epoch for self-applied PSG showed a similar trend, with sleep technologists spending 2.0s on average per epoch when scoring without recommendations, which rose to 2.1s when recommendations were introduced. Unlike traditional PSG, this effect was statistically significant (t -test $p=0.0036$, rank-sum $p=0.0039$).

Similar to the scoring accuracy, this effect was clearer when scorings were separated based on recommendation correctness and presence (see Figure 11). For both traditional and self-applied PSG, the sleep technologists spent 2.0s on average per epoch when not faced with any recommendations.

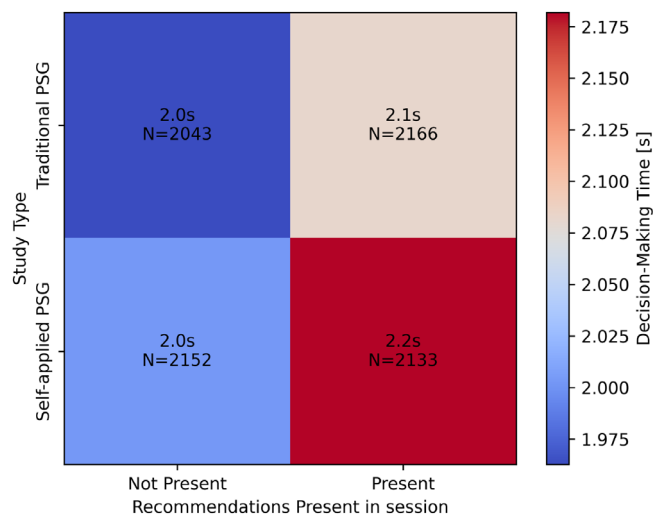


FIGURE 10 | Effect of recommendation presence on decision-making time for traditional versus self-applied PSG.

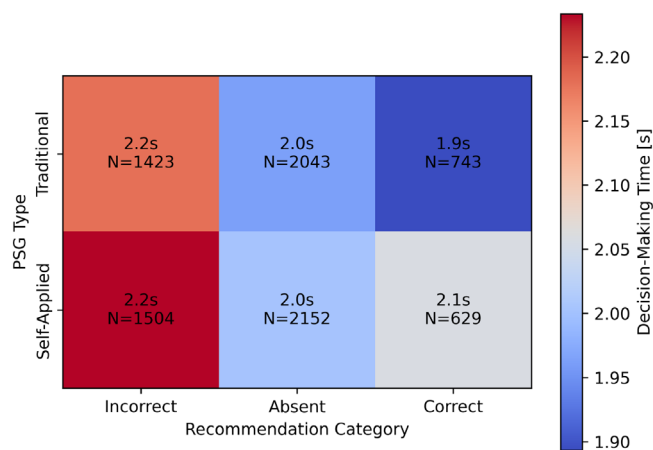


FIGURE 11 | Effect of recommendation correctness on decision-making time.

For incorrect recommendations, the sleep technologists spent 2.2s per epoch on average for both types of PSG. When recommendations are correct, however, the average time-per-epoch decreased by 0.1s for the traditional PSG; however, the decision-making time for self-applied PSG increased by 0.1s on average.

Similarly to the accuracy, ART ANOVA was used to discover which features of the recommendations and their interactions affected the decision-making time. The results of the ANOVA can be seen in Table 2, where the significance levels are marked with one to three asterisks, depending on the significance level. The effect of the presentation alone was insignificant for the decision-making time, and its interaction with the correctness and PSG type did not reach statistical significance. PSG type was found to affect the decision-making time with statistical significance, and its interaction with the correctness of the recommendations had a borderline significant effect on the decision-making time. The correctness was highly statistically significant for the decision-making time. The three-way interaction of the variables had a highly statistically significant effect.

TABLE 2 | ART ANOVA results for presentation, study type and correctness on average decision-making time.

Effect	Df	Df.res	F value	Pr(>F)
Presentation	1	2150	0.257	0.612
PSG type	1	2150	6.860	0.008**
Correctness	1	2150	38.702	5.915e-10***
Presentation: PSG Type	1	2150	4.361	0.036*
Presentation: Correctness	1	2150	0.837	0.360
PSG Type: Correctness	1	2150	0.391	0.531
Presentation: PSG Type: Correctness	1	2150	16.193	5.917e-05***

Note: Statistical significance is marked with an asterisk in the final column.

The three-way interaction between study type, recommendation presentation and recommendation correctness revealed distinct trends in decision-making time (Figure 12). For traditional PSG, for epochs featuring correct human recommendations, the sleep technologists spent an average of 1.99s per epoch and 1.85s for epochs featuring AI recommendations, or approximately 0.04s shorter when the recommendations were presented as being from AI. For self-applied PSG, correct recommendations displayed a similar trend of sleep technologists taking less time on average to score epochs with AI recommendations (2.0s) vs. human recommendations (2.3s).

Incorrect recommendations showed considerable increases in average decision-making time per epoch, as stated earlier, with sleep technologists spending on average 2.9s scoring epochs with an incorrect human recommendation versus 2.2s for an incorrect AI recommendation. Self-applied PSG reverses this trend; the sleep technologists spent 2.5s on incorrect human recommendation epochs versus 3.0s on incorrect AI recommendations.

4 | Discussion

The main contributions of this work are threefold:

1. We found no significant difference in the scoring accuracy between traditional and self-applied PSG.
2. We found that correct recommendations increased the scoring accuracy for both the traditional and self-applied PSG up to approximately 90% accuracy.
3. We found no evidence for bias toward AI recommendations over human recommendations when scoring sleep stages.

The findings of this study provide strong evidence for the potential of AI and DSS to enhance sleep stage scoring by improving accuracy and reducing decision-making time. This work contributes significantly to integrating AI-driven and

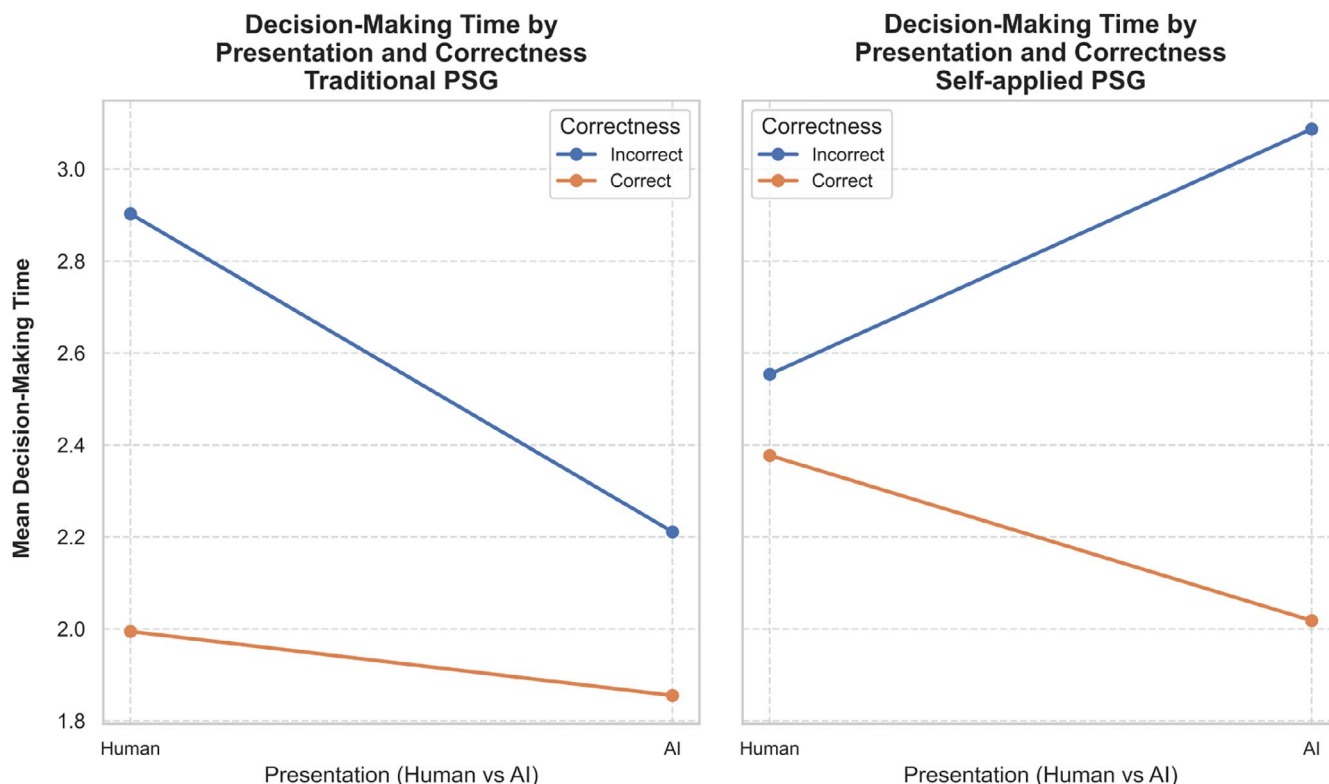


FIGURE 12 | Three-way line plot with grouped comparisons between the effect of study type, recommendation presentation and recommendation correctness on decision-making time.

automated scoring systems into the workflows of sleep technologists, paving the way for faster processes and more precise diagnostics.

4.1 | Traditional Versus Self-Applied PSG

When analysed, no evidence was found that sleep staging accuracy differed for traditional or self-applied PSGs. While the participants were slightly more accurate when scoring traditional PSG epochs, a likely explanation for this is that the majority of sleep technologists participating had not scored a self-applied PSG until this study. Along with the scoring rules for self-applied PSGs being less defined than for traditional PSGs, those two factors are the most likely to affect the scoring accuracy. Our findings align with Rusanen et al. (2023), demonstrating that self-applied PSGs can be reliably scored without additional time or accuracy penalties compared to traditional PSGs. Furthermore, we did not find that the baseline decision-making time differed between the study types, suggesting that the time taken to score self-applied PSGs is not greater than for traditional PSGs.

4.2 | Scoring Accuracy

The results showed that the baseline scoring accuracy without recommendations was in line with the work of Nikkonen et al. (2024), as sleep technologists achieved a comparable baseline agreement with the prior work exploring the inter-scorer variability in sleep staging. Sleep technologists were more likely to score self-applied PSG epochs incorrectly when presented

with incorrect recommendations than with traditional PSG epochs. Both types of PSGs showed reduced scoring accuracy with incorrect recommendations; however, self-applied PSGs suffered a significantly greater decrease relative to its baseline accuracy, with a 7.26% point drop compared to a 3.26% point drop for traditional PSGs. This discrepancy is potentially due to sleep technologists being generally less familiar with self-applied PSG signals, thus deferring the scoring decision to the recommendations. While Rusanen et al. (2023) found that self-applied PSGs tends to suffer from noisier signals than the traditional PSGs or electrode placement issues, this does not apply to this study since the scoring session periods were chosen for their signal quality.

Whether recommendations were presented as being from humans or AI had seemingly no statistically significant effect on the accuracy, indicating a high clinical acquiescence. This finding underscores the potential for seamless integration of AI tools in clinical workflows.

These results have significant clinical implementations, as both the traditional and self-applied PSG scoring show significant improvements in terms of accuracy when correct recommendations are integrated into the scoring process.

4.3 | Decision-Making Time

As Table 2 shows, when epochs with recommendations are analysed, the study type becomes a significant factor for decision time, with sleep technologists spending an average of 0.2 s less time scoring traditional PSG epochs. The correctness was also

found to be significant for the decision-making time, with sleep technologists spending 0.3 s longer scoring epochs with incorrect recommendations for the traditional PSGs and 0.1 s longer for incorrect self-applied PSG recommendations.

The interaction between study type and recommendation presentation also influenced the decision-making time, with sleep technologists spending considerably more time evaluating incorrect AI recommendations for self-applied PSGs than similar recommendations for traditional PSGs. This disparity suggests that, due to their familiarity with traditional PSGs, sleep technologists could more swiftly dismiss incorrect AI recommendations while dedicating additional effort to understanding incorrect human recommendations. Conversely, sleep technologists scored incorrect human recommendations for self-applied PSGs faster than AI recommendations. This indicates that while technologists were more inclined to defer to human recommendations, they subjected AI recommendations to greater scrutiny, reflecting a nuanced approach to clinical acquiescence in the context of novel applications.

The clinical implications for these findings are significant as the reduction in time observed in our results was 0.3 s for the traditional PSG epoch, or approximately 13% of the mean decision-making time, meaning that if an eight-hour PSG is scored in 2 h, with recommendations, the scoring time will become approximately 103 min or a 17 min gain in time. However, this gain in scoring time is not observed to such a degree for the self-applied PSG epochs, with the time reduction being 0.1 s on average, resulting in a 4% speed up, resulting in a scoring time of 114 min or a gain of 5 min.

4.4 | Study Limitations

Despite the promising contributions, the study has several notable limitations. First, the data collection period was restricted to 4 months due to the timeline of the overarching project; a longer period may have enabled the recruitment of a larger cohort of sleep technologists. Second, although the participant pool of 16 sleep technologists provided valuable insights, a larger sample size would be required to ensure statistical power and capture a more comprehensive range of inter-individual variability. While the diversity of participants is a strength, the relatively small number limits the generalisability of some conclusions.

As noted in the results, the participating sleep technologists came from diverse professional backgrounds, introducing a high level of heterogeneity that is not directly addressed in this study. Controlling for this variable would require a larger sample of participants, as the systematic impact of scorer heterogeneity on outcomes remains insufficiently understood.

On a separate note, although this study explores a method for presenting recommendations to sleep technologists, it does not include confidence scores or validity metrics alongside the recommendations. While such information could enhance explainability and promote trust in DSSs, it was not implemented here, as all recommendations were sourced from human experts and merely labelled as AI-generated, rather than produced by an automated AI scoring system.

Finally, this study focused on a particularly clean 1-h segment of PSG data to limit scope and control for the confounding effects of noise or difficult-to-score regions. As a result, the findings are not generalisable to full-night recordings or to noisier, more challenging signal data, which may elicit different responses from sleep experts and potentially alter the impact of recommendations.

5 | Conclusion

This study demonstrates that DSSs have significant potential to affect the scoring accuracy and speed of sleep technologists positively. However, while correct recommendations can make the scoring process more accurate and time-efficient, incorrect recommendations will likely have the opposite effect. Our findings emphasise the critical need for the reliability and correctness of the systems integrated into the workflows of sleep technologists. Additionally, this study provides valuable insights for further research into DSSs and implementing human-in-the-loop software that incorporates AI into sleep medicine.

Future research should focus on expanding the integration of AI in sleep diagnostics, exploring its application across diverse scoring tasks, and developing systems that empower human experts to deliver more accurate and reliable patient care in less time than currently required. The insights gained from this study pave the way for AI-driven innovations that could revolutionise sleep medicine and enhance patient outcomes. Future research should also examine the long-term effects of AI integration in the scoring process, specifically whether prolonged exposure to AI influences sleep technologists' scoring behaviours or decision-making habits. Future research could involve longer scoring periods or multiple shorter sessions to address the current limitation of insufficient sleep stage variety, incorporating more diverse and representative sleep segments. Furthermore, given time to adjust and learn to recognise the scoring patterns of the recommendations, the positive effect observed for both scoring accuracy and decision-making time could theoretically increase; however, that requires study to confirm. Research into hypodensity can also be conducted since platforms such as MicroNyx allow remote multi-scorer consensus research to be conducted more easily than previously available. Finally, incorporating information on AI uncertainty can increase the DSS's transparency, increasing the trust and clinical acquiescence of the automatic assistance provided to sleep technologists.

This work should be viewed as a preliminary—or proof of concept—analysis, both due to the limitations discussed and the range of future research directions it opens. Follow-up studies conducted at a larger scale and over extended data collection periods—particularly those designed to address the limitations identified here—would not only strengthen the current findings but also yield insights critical to the development of AI systems in sleep research and clinical practice.

This study underscores the need for the accuracy and reliability of DSS tools when incorporated into the scoring process, as incorrect recommendations were shown to impact sleep technologists' accuracy and decision-making time negatively. Furthermore, the results gathered in this study suggest that take-home PSG

solutions do not suffer from reduced accuracy in sleep scoring compared to traditional PSG. However, our findings underscore the risks of over-reliance on AI recommendations, particularly in self-applied PSG when incorrect recommendations are involved. If not carefully managed, such reliance can result in systematic errors, potentially compromising diagnostic reliability. To address these challenges, regular calibration and retraining of AI systems and enhanced training for sleep technologists to effectively collaborate with AI are crucial for mitigating risks and ensuring balanced, reliable outcomes.

Author Contributions

Benedikt Holm: conceptualization, investigation, writing – original draft, methodology, validation, visualization, writing – review and editing, software, formal analysis, project administration, data curation, supervision, resources. **Arnar Óskarsson:** software, investigation. **Björn Elvar Þorleifsson:** software, investigation. **Hörður Þór Hafsteinsson:** software, investigation. **Sigríður Sigurðardóttir:** data curation, methodology, validation, writing – review and editing. **Heiður Grétarsdóttir:** validation. **Kenan Hoelke:** validation, writing – review and editing. **Gabriel Marc Marie Jouan:** writing – original draft, writing – review and editing. **Thomas Penzel:** funding acquisition, resources, supervision. **Erna Sif Arnardóttir:** conceptualization, formal analysis, funding acquisition, methodology, project administration, resources, supervision. **María Óskarsdóttir:** conceptualization, methodology, formal analysis, funding acquisition, project administration, resources, supervision.

Acknowledgements

This work would have been impossible to accomplish without the invaluable support of Thomas Penzel and the resident sleep technologists at Charité University Hospital in Berlin. The European Society of Sleep Technologists and its president, Carlos Teixeira, receive special gratitude for allowing us to advertise the study in an email to its members. During the preparation of this article, generative AI tools, specifically ChatGPT (chatgpt.com), were employed to proofread the text and provide suggestions for improvement. The ChatGPT tool was used solely as a refinement aid and was not involved in drafting any part of this work. Additionally, AI-powered search engines, including Perplexity.ai and Semantic Scholar, were utilised for literature research and sourcing references for this article. Finally, grammatical corrections were performed using Grammarly (grammarly.com), provided through an educational account by Reykjavík University.

Conflicts of Interest

Dr. Erna Sif Arnardóttir reports personal fees from Nox Medical, Jazz Pharmaceuticals, Linde Healthcare, Wink Sleep, Apnimed, Vistor and ResMed. She is a member of medical advisory boards for Philips Sleep Medicine & Innovation Medical Board and Lilly. The other authors declare no conflicts of interest.

Data Availability Statement

The datasets used to do the work of this article are not publicly available because they contain medical data and other confidential polysomnographic data. Requests to access the datasets should be directed to sleepsupport@ru.is.

References

Adams, R. J., S. L. Appleton, A. W. Taylor, et al. 2017. "Sleep Health of Australian Adults in 2016: Results of the 2016 Sleep Health Foundation National Survey." *Sleep Health* 3, no. 1: 35–42.

American Academy of Sleep Medicine. 2023. "The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 3 Edition." <https://aasm.org/clinical-resources/scoring-manual/>.

Arnardóttir, E. S., E. Björnsdóttir, K. A. Ólafsdóttir, B. Benediksdóttir, and T. Gíslason. 2016. "Obstructive Sleep Apnoea in the General Population: Highly Prevalent but Minimal Symptoms." *European Respiratory Journal* 47, no. 1: 194–202.

Arnardóttir, E. S., A. S. Islind, M. Óskarsdóttir, et al. 2022. "The Sleep Revolution Project: The Concept and Objectives." *Journal of Sleep Research* 31, no. 4: e13630.

Asan, O., A. E. Bayrak, and A. Choudhury. 2020. "Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians." *Journal of Medical Internet Research* 22, no. 6: e15154.

Ferretti, D., A. S. Islind, K. A. Ólafsdóttir, S. Sigurðardóttir, and E. S. Arnardóttir. 2022. "The Use and Quality of 3 Nights Self-Applied Home Sleep Studies." *Sleep Medicine* 100: S305.

Garg, A. X., N. K. J. Adhikari, H. McDonald, et al. 2005. "Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review." *JAMA* 293, no. 10: 1223–1238.

Harbarth, L., E. Gößwein, D. Bodemer, and L. Schnaubert. 2024. "(Over) trusting AI Recommendations: How System and Person Variables Affect Dimensions of Complacency." *International Journal of Human-Computer Interaction* 41, no. 1: 391–410.

Holm, B. 2023. "Micronyx.is, Platform Used for Experimental Purposes." <https://micronyx.is>.

Holm, B., G. Jouan, E. Hardarson, et al. 2024. "An Optimized Framework for Processing Multicentric Polysomnographic Data Incorporating Expert Human Oversight." *Frontiers in Neuroinformatics* 18. <https://doi.org/10.3389/fninf.2024.1379932>.

Kay, M., L. A. Elkin, J. J. Higgins, and J. O. Wobbrock. 2025. "ARTool: Aligned Rank Transform for Nonparametric Factorial ANOVAs." <https://github.com/mjskay/ARTool>.

Mann, H. B., and D. R. Whitney. 1947. "On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other." *Annals of Mathematical Statistics* 18, no. 1: 50–60. <https://doi.org/10.1214/aoms/1177730491>.

McKinney, W. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, vol. 445, 51–56.

Moridian, P., A. Shoeibi, M. Khodatars, et al. 2022. "Automatic Diagnosis of Sleep Apnea From Biomedical Signals Using Artificial Intelligence Techniques: Methods, Challenges, and Future Works." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, no. 6: e1478.

Mostafa, S. S., F. Mendonça, G. Ravelo-García, and F. Morgado-Dias. 2019. "A Systematic Review of Detecting Sleep Apnea Using Deep Learning." *Sensors* 19, no. 22: 4934. <https://doi.org/10.3390/s19224934>.

Nikkonen, S., P. Somaskandhan, H. Korkalainen, et al. 2024. "Multicentre Sleep-Stage Scoring Agreement in the Sleep Revolution Project." *Journal of Sleep Research* 33, no. 1: e13956.

Parasuraman, R., and D. H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors* 52, no. 3: 381–410.

Pitkänen, M., H. Pitkänen, R. K. Nath, et al. 2024. "Temporal and Sleep Stage-Dependent Agreement in Manual Scoring of Respiratory Events." *Journal of Sleep Research* 34: e14391.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria." <https://www.R-project.org/>.

Rayan, A., A. B. Szabo, and L. Genzel. 2023. "The Pros and Cons of Using Automated Sleep Scoring in Sleep Research: Comparative Analysis of Automated Sleep Scoring in Human and Rodents: Advantages and Limitations." *Sleep* 47, no. 1: zsad275. <https://doi.org/10.1093/sleep/zsad275>.

Rosenberg, R. S., and S. Van Hout. 2014. "The American Academy of Sleep Medicine Inter-Scorer Reliability Program: Respiratory Events." *Journal of Clinical Sleep Medicine* 10, no. 4: 447–454.

Rusanen, M., H. Korkalainen, H. Gretarsdottir, et al. 2024. "Self-Applied Somnography: Technical Feasibility of Electroencephalography and Electro-Oculography Signal Characteristics in Sleep Staging of Suspected Sleep-Disordered Adults." *Journal of Sleep Research* 33, no. 2: e13977.

Rusanen, M., H. Korkalainen, H. Grétarsdóttir, et al. 2023. "Self-Applied Somnography: Technical Feasibility of Electroencephalography and Electro-Oculography Signal Characteristics in Sleep Staging of Suspected Sleep-Disordered Adults." *Journal of Sleep Research* 33: e13977. <https://doi.org/10.1111/jsr.13977>.

Sprague, R. H. 1980. "A Framework for the Development of Decision Support Systems." *MIS Quarterly* 4, no. 4: 1–26. <http://www.jstor.org/stable/248957>.

Varsha, P. S. 2023. "How Can We Manage Biases in Artificial Intelligence Systems—A Systematic Literature Review." *International Journal of Information Management Data Insights* 3, no. 1: 100165. <https://doi.org/10.1016/j.jjime.2023.100165>.

Virtanen, P., R. Gommers, T. E. Oliphant, et al. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17, no. 3: 261–272. <https://doi.org/10.1038/s41592-019-0686-2>.

Wara, T. U., A. H. Fahad, A. S. Das, and M. M. H. Shawon. 2024. "A Systematic Review on Sleep Stage Classification and Sleep Disorder Detection Using Artificial Intelligence." arXiv [cs.LG].

Zeng, L. N., Q. Q. Zong, Y. Yang, et al. 2020. "Gender Difference in the Prevalence of Insomnia: A Meta-Analysis of Observational Studies." *Frontiers in Psychiatry* 11: 577429. <https://doi.org/10.3389/fpsy.2020.577429>.

Appendix A

Study Completion Questionnaire

Participant Information

1. What is your name?

(We will not use your personal identity for any analysis or publication, and it will remain completely confidential.)

Answer: _____.

2. What is your email?

(If this is not prefilled, please use the same email you used to log into MicroNyx.)

Answer: _____.

3. Where are you working?

(Center, City, Country)

Answer: _____.

Study Awareness

1. I heard about this study from:
 - The invitation email sent to the Sleep Revolution.
 - The talk held by Dr. Erna Sif Arnardóttir at the Sleep Europe conference.
 - The talk held by Dr. Erna Sif Arnardóttir at the ERS conference.
 - Email sent to the members of the ESST.
 - Other: _____.

2. Are you part of the Sleep Revolution?

- Yes
- No

User Experience

This section of the questionnaire is designed to gain insight into the ease of use of the MicroNyx platform.

1. The system was easy to use.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree.

2. I felt that I could reliably read, interpret and score the signals in the interface.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree.

3. The scoring recommendations were easy to see.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree.

4. It was clear which scoring recommendations were from a human and which were from an AI.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree.

5. I felt the recommendations were helpful when scoring.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree.

Additional Feedback

If you have any further comments, please write them here. We are very happy to hear your feedback.

Answer: _____.