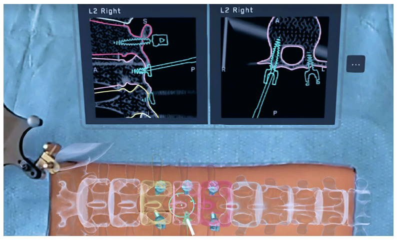

# Cognitive Workload, Usability, and Operator Preference in Augmented Reality Spinal Navigation

Marharyta Krylova, MBE[1]; Heeya Shah[1]; Florentin Liebmann, MS, PhD[1]

[1]Medivis Inc., New York City, NY

## ABSTRACT

**Background.** Pedicle screw accuracy has converged across contemporary navigation, robotic, and augmented reality (AR) platforms. In most cases, the differences that remain between systems are in workflow and in the burden placed on the operator rather than in whether a screw lands within the pedicle. Conventional navigation presents orthogonal two dimensional slices on a monitor positioned away from the operative field, obliging the surgeon to look away from the patient and to mentally reconstruct a three dimensional trajectory before mapping it back onto the anatomy in front of them.

**Objective.** To compare perceived workload, perceived usability, and operator preference between conventional two dimensional slice navigation display and AR three-dimensional on-body projection during minimally invasive pedicle screw placement.

**Methods.** A preclinical cadaveric study was conducted using a within operator paired design. A Medivis Spine Navigation system was used in two display conditions on matched vertebra pairs. The left pedicle was instrumented with a workstation monitor showing two dimensional X and Z slices only (Screen) and the right pedicle with a headset projecting a three dimensional spine model and the planned screw trajectory onto the specimen alongside a floating two dimensional slice panel, identical to the workstation display in the Screen condition (AR). Each operator was allocated 4 vertebral levels and 8 screws, 4 under each condition, and completed the NASA-TLX, the System Usability Scale, and preference questions on completion. Mental demand from 0 to 100 was rated by the operator immediately after each screw placement. Gaze location was classified from recorded video by a computer vision-based classifier. One of four enrolled operators was excluded in full for a protocol deviation, leaving three operators for analysis. All analyses are descriptive and no inferential testing was performed due to a limited sample size.

**Results.** NASA-TLX favored AR on five of six subscales. Raw TLX was 18.06 under AR versus 23.78 under Screen, 24.1% difference. Effort showed the largest reduction (26.67 versus 41.00, 35.0% lower), followed by Mental demand (28.33 versus 36.33), Frustration (9.67 versus 14.33), Physical demand (23.67 versus 27.67), and Temporal demand (11.00 versus 14.67). Performance was unchanged (9.00 versus 8.67). Mental demand rated after each placement was lower under AR for all three operators, with an operator level mean of 27.78 versus 39.44 and within operator reductions of 3.33, 15.00, and 16.67 points. System Usability Scale composites were 91.67 (±1.44) for AR and 95.00 (±2.50) for Screen; the 3.33 point difference in favor of the Screen condition came from four of the ten items (Q4, Q7, Q8, and Q10). All three operators preferred AR overall, found it easier to set the trajectory, and reported feeling faster; two of three felt more confident and one of three felt more accurate. Under the Screen condition, 70.59% of 25,095 analyzed frames were directed at the navigation display and 29.41% at the operating field.

**Conclusions.** In this cadaveric feasibility study, AR three-dimensional on-body projection was associated with lower reported workload and unanimous operator preference relative to two dimensional slice display, while self assessed performance was unchanged. Screen scored marginally higher on usability, and that difference was limited to items on learning and required technical support. The findings are descriptive, rest on three operators, and require confirmation in a larger cohort and in a live surgical environment.

**Keywords:** augmented reality; surgical navigation; pedicle screw; minimally invasive spine surgery; cognitive load; NASA-TLX; System Usability Scale; surgical navigation preference; cadaveric study

## INTRODUCTION

Computer assisted navigation, robotic assistance, and augmented reality (AR) all now place pedicle screws with consistently high accuracy [1–4]. A systematic review of AR assisted placement reported that 2,095 of 2,232 screws, or 93.1%, were judged clinically accurate across cadaveric, phantom, and clinical series [5]. Because accuracy no longer reliably discriminates between approaches, its value as a comparative endpoint has diminished, and attention has moved toward how a system behaves in the hands of the surgeon who must use it.

Two properties of conventional navigation are relevant to that question. First, the anatomy is rendered on a monitor positioned away from the operative field, so the operator must repeatedly break visual contact with the operative field in order to consult it. Second, that monitor presents orthogonal two dimensional slices, so the operator must mentally assemble a three-dimensional trajectory from separate planes and then transpose it onto the patient in front of them. Robotic platforms provide strong preoperative planning but are driven largely from a separate console and screen, and their advantages have not clearly extended to reducing the cognitive load of three-dimensional trajectory reconstruction during execution [6].

AR responds to both properties by rendering the three-dimensional anatomy and the planned trajectory within the operator's own view of the surgical field. If operated as intended, the most direct consequence should be a reduction in the cognitive workload the operator carries while executing a planned trajectory. This study was designed to test that proposition under controlled cadaveric conditions using Medivis Spine Navigation [7, 8]. This report covers the perceived workload, perceived usability, modality preference, and gaze distribution. Nothing in this report should be read as a claim about the speed or the accuracy of screw placement in either condition.

## METHODS

### Study design and setting

This was a preclinical cadaveric quality assessment and improvement study employing a within operator paired design conducted in a single cadaveric laboratory. One navigation system was used in two display conditions. Each vertebra served as its own matched pair and was instrumented on both sides under the same registration, with side fixed to condition: the left pedicle was placed under the Screen condition and the right pedicle under the AR condition. All placements were minimally invasive and percutaneous. Because each operator served as their own control and both sides of a vertebra were placed under a single registration, bone quality, level difficulty, pedicle dimensions, and registration were held constant within every pair.

**Display conditions**

The two conditions are defined in Table 1. Both were radiation free during placement. The AR condition supplied three dimensional spatial context in addition to the same slice data available in the Screen condition. Any difference observed between conditions therefore reflects the AR system as deployed and not display position in isolation.

**Table 1.** Definition of the two display conditions.

| Condition | Display provided to the operator |
|---|---|
| Screen | Two dimensional X/Z image slices on the workstation monitor only. No three-dimensional rendering and no projection onto the specimen. Represents conventional two dimensional slice navigation display. |
| AR | A headset projecting a three dimensional segmentation of the spine with the planned screw trajectory onto the specimen, together with a floating panel of two dimensional X/Z slices positioned above the surgical field. |

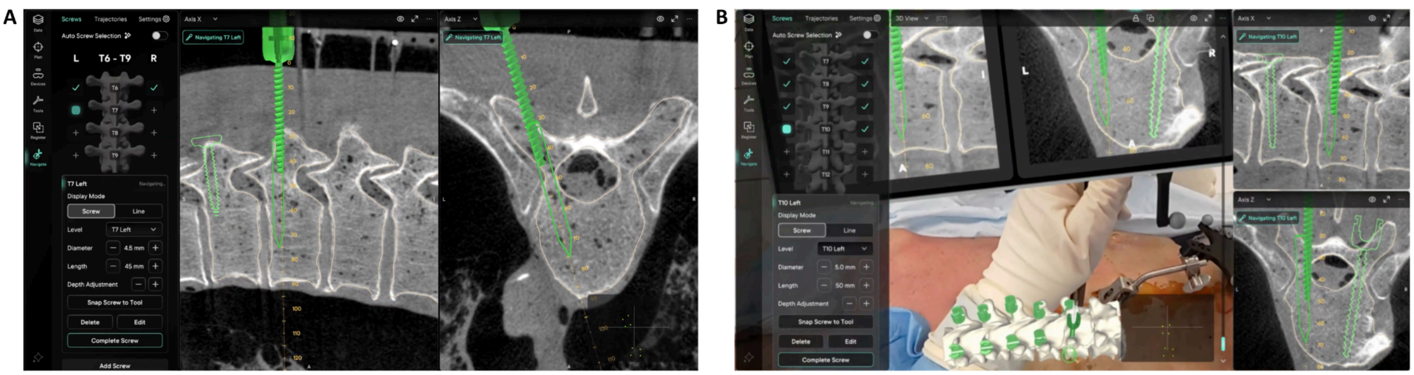


**Figure 1.** Two display conditions used throughout the cadaver lab. A) Two-dimensional X/Z screen condition displayed on the workstation monitor only; B) Feed from the headset with the augmented reality three-dimensional segmentation of the spine alongside a floating panel with X/Z slices.

**Specimens, levels, and instrumentation**

The study was designed around two human cadavers and 16 thoracolumbar matched within vertebra pairs, for a planned total of 32 study screws. Each operator was allocated 4 thoracic or lumbar pairs. Screw diameter and length were fixed by level in advance of the laboratory. Inclusion required cadavers with intact, instrumentable anatomy and adequate planning imaging. No specimens had prior hardware, fracture, deformity, or destruction that would preclude placement.

**Operators**

Four operators, comprising two experienced spine surgeons and two residents, participated. Every operator performed both display conditions. Operators were not alternated: each completed their full block of 8 screws before the next operator began. Such a decision was driven by the length of the workflow.

**Experimental procedure**

The workstation was configured to display two dimensional slices only for the Screen condition. The headset was confirmed charged and functional with a spare unit on standby. All tracked instruments were verified against the system, and the recording camera was tested to confirm that it captured the operator's gaze in full and allowed reliable distinction between the workstation monitor and the operative field. A tracked reference array was affixed. A three-dimensional spin covering the segment was acquired on a mobile imaging system (GE OEC; Boston, MA) and imported to the workstation. The scan was registered to the specimen using the reference array, after which the operator qualitatively verified registration accuracy before the segment was declared live. For each pedicle, the navigation technician set the display for that side and announced the level, side, and condition aloud. The operator defined the planned trajectory under the assigned display condition and the trajectory was saved. The trajectory was fixed at that point and was not adjusted during placement. The opening was then drilled and tapped following a standard clinical workflow, and the screw was placed to full seating. Immediately after completing each screw, the operator announced a mental demand rating from 0 to 100 for that placement, which the data recorder logged together with any additional technical events. An accuracy recheck closed each segment and the coordinator confirmed that both sides of every pair were complete before the segment was marked closed. On completing all 8 of their allocated screws, each operator was administered the full anonymized survey via an online platform, Qualtrics [9].

**Outcome measures**

The outcome measures reported in this report are set out in Table 2. On all six NASA-TLX subscales as reported here, a lower score indicates a lower reported burden [10]. This includes the Performance subscale, which is reverse oriented on the instrument such that a low score denotes better self assessed performance. On the System Usability Scale, a higher composite score indicates better perceived usability [11].

**Table 2.** Outcome measures reported in this analysis.

| Domain | Instrument | Administration and scoring |
|---|---|---|
| Perceived workload | NASA-TLX [10] | Six subscales scored 0 to 100. Administered once per operator per condition through an electronic survey on completion of the 8 screw block. |
| Perceived workload | Mental demand per placement | Single 0 to 100 rating announced verbally by the operator immediately after each screw placement and logged by the data recorder. |
| Perceived usability | System Usability Scale [11] | Ten items scored 1 to 5, administered once per operator per condition. Scored conventionally: odd numbered items contribute the response minus one, even numbered items contribute five minus the response, and the sum is multiplied by 2.5 to yield a composite from 0 to 100. |
| Modality preference | Forced choice items | Operators selected one modality overall and answered separate items on ease of defining the trajectory, confidence, perceived speed, and perceived accuracy. |
| Attention allocation | Gaze classification | Proportion of analyzed video frames in which the operator's focus was directed at the operating field versus the two dimensional navigation display, classified automatically from the recorded video, using a computer vision algorithm. |

**Protocol deviation and exclusion**

Three of the four operators are included in this analysis: two experienced spine surgeons (O1 and O3) and one resident (O4). Each completed both display conditions and returned complete NASA-TLX, System Usability Scale, and preference data. Operator O2 was excluded in full for a protocol deviation. Individual screws were also excluded for deviations that altered the workflow. O3 introduced additional drilling and tapping attempts on both screws at one level, and O4 did so on two screws in the Screen condition, so the adjustment affected three screws in the Screen condition and one in the AR condition. Separately, one Screen recording for O1 failed. Where only one side of a vertebra was affected, the matched screw on the contralateral side was also excluded to preserve the paired design. Four matched pairs were excluded in total, one each for O1 and O3 and two for O4, leaving three placements per condition for O1 and O3 and two per condition for O4.

**Gaze direction processing**

Gaze direction was determined from head orientation in the recorded video using MediaPipe Face Landmarker [12]. Videos were sampled at 10 frames per second. For each camera setup, representative frames in which the operator was oriented toward the navigation display or operative field were used for calibration. Facial landmarks were used to estimate head pitch and yaw, and the median values for each target served as reference points. Each analyzed frame was classified according to the nearest reference point by Euclidean distance. Frames that could not be assigned to a calibrated gaze location were manually reviewed and annotated. Frames in which no face was detected were classified as Operative Field, as the operator's face was often obscured by the camera angle when looking down toward the surgical field. Classification was rule-based and did not involve training a supervised model. The proportion of time directed toward each location was calculated from the proportion of analyzed frames assigned to that location.

**Statistical analysis**

All summary data and figures were generated using a Python 3.12 pipeline [13]. *pandas* 3.0.2 and *numpy* 2.4.4 libraries were used to load, clean, and aggregate the gaze, NASA-TLX, System Usability Scale, and preference data and to compute the descriptive statistics reported [14, 15]. Figures 1 through 5 were produced with *matplotlib* 3.10.8 [16].

All results reported are descriptive. With three operators each contributing a single observation per condition on the survey instruments, no inferential test would be interpretable, and none was performed. No probability values are reported and none should be inferred. Continuous measures are summarized as means with standard deviations. Preference results are reported as counts out of three alongside the corresponding percentage.

**Ethics and consent**

Printed participation consent was provided to each operator on the day of the laboratory and was completed before that operator began. All operators whose data are included in this analysis signed consent. Video recordings for gaze analysis and all raw data are stored in a deidentified format.

## RESULTS

**Perceived workload (NASA-TLX)**

AR scored lower than Screen on five of the six subscales (Table 3, Figure 2). Effort showed the largest reduction (26.67 versus 41.00, 14.33 points, 35.0% lower), followed by Mental demand (28.33 versus 36.33, 8.00 points, 22.0%), Frustration (9.67 versus 14.33, 4.66 points, 32.5%), Physical demand (23.67 versus 27.67, 4.00 points, 14.5%), and Temporal demand (11.00 versus 14.67, 3.67 points, 25.0%). Performance was 9.00 under AR and 8.67 under Screen, a difference of 0.33 points. Total raw TLX was 18.06 under AR and 23.78 under Screen, 5.72 points or 24.1% lower. Standard deviations were wide relative to these differences and the two conditions overlapped on every subscale.

**Table 3.** NASA-TLX subscale scores by display condition. Three operators, one rating per operator per condition. A lower score indicates a lower reported burden on every subscale shown, including Performance, which is reverse oriented on the instrument. Raw TLX is the unweighted mean of the six subscale means. Values are descriptive; no inferential testing was performed and standard deviations overlap substantially between conditions on all subscales.

| Subscale | AR, mean (SD) | Screen, mean (SD) | Difference | Relative reduction |
|---|---|---|---|---|
| Mental demand | 28.33 (±22.55) | 36.33 (±15.18) | 8.00 lower under AR | 22.0% |
| Physical demand | 23.67 (±7.77) | 27.67 (±12.50) | 4.00 lower under AR | 14.5% |
| Temporal demand | 11.00 (±1.73) | 14.67 (±5.03) | 3.67 lower under AR | 25.0% |
| Performance | 9.00 (±15.59) | 8.67 (±15.01) | 0.33 higher under AR | 3.8% |
| Effort | 26.67 (±20.82) | 41.00 (±18.19) | 14.33 lower under AR | 35.0% |
| Frustration | 9.67 (±9.50) | 14.33 (±14.01) | 4.66 lower under AR | 32.5% |
| **Raw TLX** | **18.06** | **23.78** | **5.72 lower under AR** | **24.1%** |

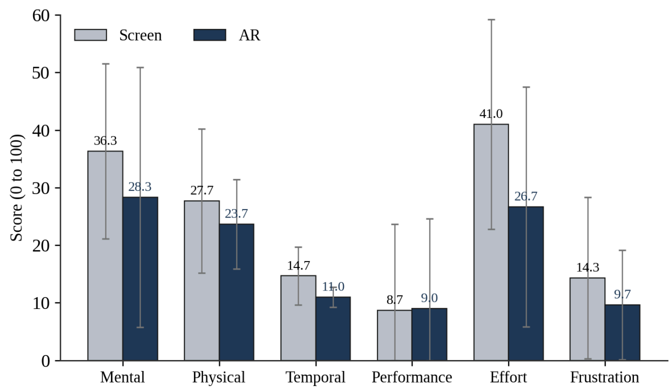


**Figure 2.** NASA-TLX subscale scores by display condition. Bars are the mean of three operators and error bars represent one standard deviation, truncated at zero where the standard deviation exceeds the mean. A lower score indicates a lower reported burden on every subscale, including Performance, which is reverse oriented on the instrument.

**Mental demand rated after each placement**

Mental demand was lower under AR for all three operators (Table 4, Figure 3). The operator level mean was 27.78 (±20.57) under AR and 39.44 (±16.86) under Screen. Reductions were 16.67 points for O1 (33.33 versus 50.00, 33.3%), 3.33 points for O3 (45.00 versus 48.33, 6.9%), and 15.00 points for O4 (5.00 versus 20.00, 75.0%).

**Table 4.** Mental demand rated by the operator immediately after each screw placement. Ratings were provided on a 0 to 100 scale. Standard deviations in the operator rows describe the dispersion of that operator's own ratings under that condition. The number of ratings available for analysis is smaller than the number of screws allocated to each operator due to deviation from the study protocol.

| Operator | AR, mean (SD) | N (AR) | Screen, mean (SD) | N (Screen) | Reduction under AR |
|---|---|---|---|---|---|
| O1 (surgeon) | 33.33 (±5.77) | 3 | 50.00 (±0.00) | 3 | 16.67 points, 33.3% |
| O3 (surgeon) | 45.00 (±5.00) | 3 | 48.33 (±10.41) | 3 | 3.33 points, 6.9% |
| O4 (resident) | 5.00 (±0.00) | 2 | 20.00 (±0.00) | 2 | 15.00 points, 75.0% |
| **Operator level mean** | **27.78 (20.57)** | **8 total** | **39.44 (16.86)** | **8 total** | **11.67 points on average** |

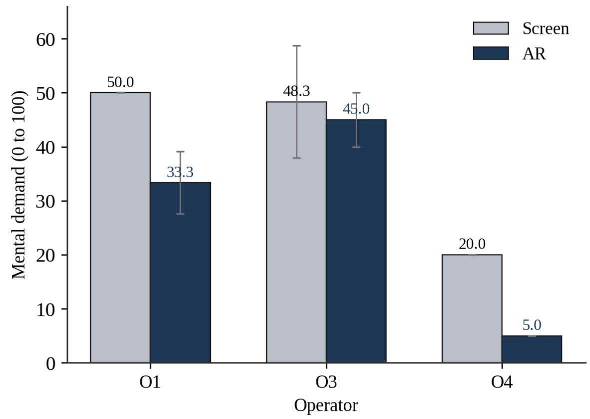


**Figure 3.** Mental demand rated by the operator immediately after each screw placement, shown by operator and display condition. Error bars represent one standard deviation of the ratings available for that operator and condition. Every operator rated the AR condition lower than the Screen condition.

Operator mean mental demand ratings spanned 5.00 to 45.00 under AR and 20.00 to 50.00 under Screen. O4 rated both conditions at the low end of the scale (5.00 and 20.00). Eight ratings per condition were available for analysis rather than the twelve allocated, the remainder having been excluded for protocol deviation.

**Perceived usability (System Usability Scale)**

Composite System Usability Scale scores were 95.00 (±2.50) for Screen and 91.67 (±1.44) for AR, a difference of 3.33 points (Table 5, Figure 4). Both are above 80, the level conventionally described as excellent, and well above the commonly cited average of 68 [17, 18]. Six of the ten items scored identically under both conditions (Q1, Q2, Q3, Q5, Q6, and Q9). The difference came from four items: Q4, on needing the support of a technical person (AR 3.00 versus Screen 2.67); Q7, on how quickly most people would learn the system (4.67 versus 5.00); Q8, on cumbersomeness (1.33 versus 1.00); and Q10, on how much had to be learned before getting started (1.67 versus 1.33). Q4 and Q10 form the learnability component of the scale as it is conventionally factored [19].

**Table 5.** System Usability Scale item means and composite scores by display condition. Items are scored 1 to 5. Odd numbered items are positively worded, so a higher response is favorable; even numbered items are negatively worded, so a lower response is favorable. Item themes are paraphrased from the instrument.

| Item | Item theme, paraphrased | AR | Screen | Differs |
|---|---|---|---|---|
| Q1 | Would like to use the system frequently | 5.00 | 5.00 | |
| Q2 | Found the system unnecessarily complex | 1.00 | 1.00 | |
| Q3 | Found the system easy to use | 5.00 | 5.00 | |
| Q4 | Would need the support of a technical person | 3.00 | 2.67 | Yes |
| Q5 | Found the functions well integrated | 5.00 | 5.00 | |
| Q6 | Found too much inconsistency in the system | 1.00 | 1.00 | |
| Q7 | Most people would learn to use it very quickly | 4.67 | 5.00 | Yes |
| Q8 | Found the system cumbersome to use | 1.33 | 1.00 | Yes |
| Q9 | Felt confident using the system | 5.00 | 5.00 | |
| Q10 | Needed to learn a lot before getting going | 1.67 | 1.33 | Yes |
| | **Composite score, mean (SD)** | **91.67 (±1.44)** | **95.00 (±2.50)** | |

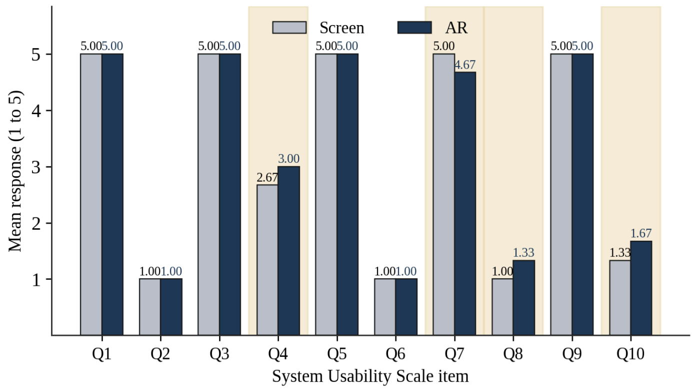


**Figure 4.** System Usability Scale item means by display condition. Shaded columns mark the only four items on which the two conditions differed. Odd numbered items are positively worded, so a higher response is favorable; even numbered items are negatively worded, so a lower response is favorable.

### Modality preference

All three operators preferred AR overall, all three found it easier to set the trajectory, and all three reported feeling faster with it (3 of 3 on each item; Table 6). Two of three reported feeling more confident under AR and one of three reported feeling more accurate.

**Table 6.** Forced choice modality preference among the three included operators. Responses are subjective impressions recorded after the operator had completed both conditions. The item addressing perceived speed reflects how placement felt to the operator and is not a measurement of elapsed time. Percentages are shown for comparability but rest on a denominator of three.

| Item | Operators favoring AR | Percentage |
|---|---|---|
| Overall preference | 3 of 3 | 100.0% |
| Easier to set the trajectory | 3 of 3 | 100.0% |
| Felt faster | 3 of 3 | 100.0% |
| Felt more confident | 2 of 3 | 66.7% |
| Felt more accurate | 1 of 3 | 33.3% |

### Gaze distribution

Gaze was classified separately for each condition (Table 7, Figure 5). Under the Screen condition, 70.59% of 25,095 analyzed frames were directed at the two dimensional navigation display and 29.41% at the operating field. The proportion on the operating field ranged from 22.49% for O4 to 42.45% for O3. Under AR, all frames were directed at the operating field.

**Table 7.** Gaze distribution by operator in the Screen condition. Percentages are the proportion of analyzed video frames classified into each category by the gaze classification algorithm. Under AR, all analyzed frames were classified to the operating field.

| Operator | Operating field | 2D navigation display | Frames analyzed |
|---|---|---|---|
| O1 | 23.54% | 76.46% | 6,063 |
| O3 | 42.45% | 57.55% | 8,382 |
| O4 | 22.49% | 77.51% | 10,650 |
| **All operators** | **29.41%** | **70.59%** | **25,095** |

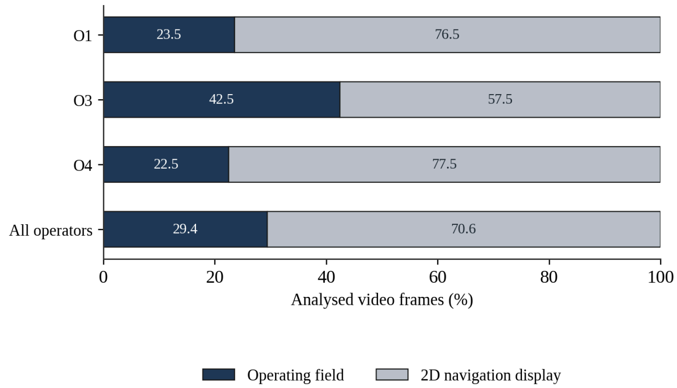


**Figure 5.** Gaze distribution by operator, expressed as the percentage of analyzed video frames within the Screen condition classified to the operating field and to the two dimensional navigation display.

## DISCUSSION

### Principal findings

The results are consistent with AR reducing the cognitive, physical, and temporal work needed to execute a planned trajectory rather than changing the outcome of that trajectory. Workload fell on the subscales describing effort and mental demand while self assessed performance did not move, which suggests operators reached a similar subjective standard at a lower cost to themselves. The gaze data offer a plausible mechanism. Under the Screen condition, operators spent most of each placement looking at a monitor away from the specimen, whereas under AR the navigation content was rendered inside their view of the field, so consulting it did not require looking away. A reduction in Effort is what would be expected if that repeated shift of attention, together with the mental reconstruction of a trajectory from separate slice planes, accounts for a meaningful part of the burden of slice based navigation. Once accuracy has converged across modalities [1, 2, 3, 4, 5], what remains is how much attention a system consumes, how much mental translation it demands, and whether surgeons want to use it. The present data point the same way on all three, but at this sample size they establish no more than that.

### Interpretation of the usability result

The usability result runs against the other findings and is better read as a familiarity and ergonomics effect. Every operator used monitor based navigation routinely and none had as extensive of an experience with the AR workflow, so the items on learning and required support are exactly where AR would be expected to score lower on early exposure. The items asking whether the system worked well were scored identically. Perceived cumbersomeness is a reasonable observation about a worn headset compared with a monitor already present in the room. Both effects are addressable, the first through structured onboarding and the second through the physical design of the device. Neither indicates that AR is harder to use once learned, and this study cannot determine whether the gap closes with training at this stage.

### Relation to existing literature

Reported accuracy for AR assisted pedicle screw placement is high and broadly comparable to that achieved with conventional navigation and robotic assistance [1–5], and AR guided placement has previously been evaluated in cadaveric models [20]. Against that background, workload and acceptability rather than accuracy are the measures on which contemporary platforms can be meaningfully separated, and the present data indicate a difference in that domain. Robotic platforms, which likewise depend on a display positioned away from the operative field, have not demonstrated a comparable reduction in intraoperative cognitive burden [6].

### Limitations

This was a feasibility quality assessment and improvement study of three operators in one laboratory over two days, so all results are descriptive and no inferential testing is appropriate. Excluding O2 removed one of the two residents and with it the planned comparison between surgeons and residents. Because AR adds three dimensional context to the same slice data, the comparison reflects the system as deployed rather than display position alone. Side was fixed to condition and operators were not alternated, confounding condition with side, handedness, time, and fatigue. The gaze classification cannot separate attention to the anatomy from attention to the panel projected over it, so the AR value describes where the operator looked rather than what they attended to. Mental demand ratings were available for 8 of the 12 allocated placements per condition. The cadaveric setting lacks bleeding and respiration and is therefore not fully representative of clinical practice. All three instruments are unblinded self-report.

### Future directions

A larger operator cohort balanced for experience level would permit the prespecified comparison between attending surgeons and residents and would support inferential analysis. Evaluation in a live surgical environment would establish whether the reduction in reported workload persists under the physiological and time pressures of clinical practice.

## CONCLUSIONS

In this cadaveric comparison of two dimensional slice navigation display against AR three-dimensional on-body projection during minimally invasive pedicle screw placement, three operators reported lower workload under AR on five of six NASA-TLX subscales and in mental demand rated after each placement, while self assessed performance was unchanged. All three preferred AR and all three found it easier to set the trajectory under the AR condition. Usability was rated excellent for both displays, with Screen marginally ahead on items concerning learning and required support. Under the Screen condition, operators spent most of each placement looking at the navigation monitor; under AR their gaze stayed on the operating field. These findings are descriptive and come from three operators in a cadaveric setting. They warrant further evaluation of AR navigation in a larger cohort and in live surgery, and they identify rigorous training and device ergonomics as the targets for improvement.

## DECLARATIONS

**Consent to participate and deidentified data use.** All data were collected for quality assurance and quality improvement purposes and are not intended to contribute to generalizable knowledge. Printed participation consent was provided to each operator on the day of the laboratory and completed before that operator began. All operators included in this analysis signed consent.

**Availability of data and materials.** All data, including video recordings used for gaze analysis, are stored in deidentified form. Deidentified data and survey materials may be made available to qualified researchers upon reasonable request, subject to review and approval by the study authors.

**Funding.** This study was funded in full by Medivis Inc. No external funding was received.

**Author contributions.** MK led study conceptualization, data analysis and visualization, and manuscript drafting and editing. HS conducted data collection and analysis as well as manuscript editing and review. FL contributed to study conceptualization, data analysis, navigation system operation, and manuscript editing and review.

**Acknowledgments.** We thank the entire Medivis team for their involvement in this work. We are grateful to Osamah Choudhry, Christopher Morley, Long Qian, Marco von Atzigen, and Erika LaCasse for their contributions to the initial study conceptualization and for their extensive support during the cadaver lab. We also recognize Hung Ngo and Omer Nossek for their assistance with data collection, and Jesse Gallant for operating the OEC C-arm during imaging.

## REFERENCES

[1] Dennler C, Jaberg L, Spirig J, et al. Augmented reality-based navigation increases precision of pedicle screw insertion. *J Orthop Surg Res* 2020; 15: 174.

[2] Felix B, Kalatar SB, Moatz B, et al. Augmented reality spine surgery navigation: Increasing pedicle screw insertion accuracy for both open and minimally invasive spine surgeries: Increasing pedicle screw insertion accuracy for both open and minimally invasive spine surgeries. *Spine (Phila Pa 1976)* 2022; 47: 865–872.

[3] Garg B, Mehta N. Great Expectations with Augmented Reality in Spine Surgery: Hope or Hype?A commentary on the article “Operator independent reliability of direct augmented reality navigated pedicle screw placement and rod bending” by Farshad et al. *N Am Spine Soc J* 2022; 10: 100117.

[4] Elmi-Terander A, Skulason H, Söderman M, et al. Surgical navigation technology based on augmented reality and integrated 3D intraoperative imaging: A spine cadaveric feasibility and accuracy study: A spine cadaveric feasibility and accuracy study. *Spine (Phila Pa 1976)* 2016; 41: E1303–E1311.

[5] Youssef S, McDonnell JM, Wilson KV, et al. Accuracy of augmented reality-assisted pedicle screw placement: a systematic review. *Eur Spine J* 2024; 33: 974–984.

[6] Hu X, Lieberman IH. What is the learning curve for robotic-assisted pedicle screw placement in spine surgery? *Clin Orthop Relat Res* 2014; 472: 1839–1844.

[7] 510(k) Premarket Notification, https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K233725 (accessed August 3, 2026).

[8] Spine Navigation, https://www.medivis.com/spine-navigation (accessed August 3, 2026).

[9] Qualtrics XM: The Leading Experience Management Software. *Qualtrics*, https://www.qualtrics.com/ (accessed August 3, 2026).

[10] Hart SG, Staveland LE. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In: Hancock PA, Meshkati N (eds) *Advances in Psychology*. Elsevier, pp. 139–183.

[11] Brooke J. SUS: a quick and dirty usability scale. In: *Usability Evaluation in Industry. London: Taylor and Francis*. Thomas B, Weerdmeester BA, McClelland IL, 1996, pp. 189–194.

[12] Face landmark detection guide. *Google for Developers*, https://developers.google.com/edge/mediapipe/solutions/vision/face_landmarker (accessed August 7, 2026).

[13] Python Release Python 3.12.0. *Python.org*, https://www.python.org/downloads/release/python-3120/ (accessed August 3, 2026).

[14] pandas. *PyPI*, https://pypi.org/project/pandas/ (accessed August 3, 2026).

[15] numpy. *PyPI*, https://pypi.org/project/numpy/ (accessed August 3, 2026).

[16] matplotlib. *PyPI*, https://pypi.org/project/matplotlib/ (accessed August 3, 2026).

[17] Sauro J. *A Practical Guide to the System Usability Scale. Denver: Measuring Usability LLC*. 2011.

[18] Bangor A, Kortum P, Miller JT. Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies archive* 2009; 4: 114–123.

[19] Lewis JR, Sauro J. The factor structure of the system usability scale. In: *Human Centered Design*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 94–103.

[20] Spirig JM, Roner S, Liebmann F, et al. Augmented reality-navigated pedicle screw placement: a cadaveric pilot study. *Eur Spine J* 2021; 30: 3731–3737.