Surgical Endoscopy (2019) 33:2093–2103
https://doi.org/10.1007/s00464-018-6480-x

The Heidelberg VR Score: development and validation of a composite score for laparoscopic virtual reality training

Mona W. Schmidt[1] · Karl-Friedrich Kowalewski[1] · Marc L. Schmidt[2] · Erica Wennberg[1] · Carly R. Garrow[1] · Sang Paik[1] · Laura Benner[3] · Marlies P. Schijven[4] · Beat P. Müller-Stich[1] · Felix Nickel[1]

Received: 3 September 2018 / Accepted: 11 October 2018 / Published online: 16 October 2018

Abstract

Introduction Virtual reality (VR-)trainers are well integrated in laparoscopic surgical training. However, objective feedback is often provided in the form of single parameters, e.g., time or number of movements, making comparisons and evaluation of trainees' overall performance difficult. Therefore, a new standard for reporting outcome data is highly needed. The aim of this study was to create a weighted, expert-based composite score, to offer simple and direct evaluation of laparoscopic performance on common VR-trainers.

Materials and methods An integrated analytic hierarchy process-Delphi survey was conducted with 14 international experts to achieve a consensus on the importance of different skill categories and parameters in evaluation of laparoscopic performance. A scoring algorithm was established to allow comparability between tasks and VR-trainers. A weighted composite score was calculated for basic skills tasks and peg transfer on the LapMentor™ II and III and validated for both VR-trainers.

Results Five major skill categories (time, efficiency, safety, dexterity, and outcome) were identified and weighted in two Delphi rounds. Safety, with a weight of 67%, was determined the most important category, followed by efficiency with 17%. The LapMentor™-specific score was validated using 15 (14) novices and 9 experts; the score was able to differentiate between both groups for basic skills tasks and peg transfer (LapMentor™ II: Exp: 86.5 ± 12.7, Nov. 52.8 ± 18.3; $p < 0.001$; LapMentor™ III: Exp: 80.8 ± 7.1, Nov: 50.6 ± 16.9; $p < 0.001$).

Conclusion An effective and simple performance measurement was established to propose a new standard in analyzing and reporting VR outcome data—the Heidelberg virtual reality (VR) score. The scoring algorithm and the consensus results on the importance of different skill aspects in laparoscopic surgery are universally applicable and can be transferred to any simulator or task. By incorporating specific expert baseline data for the respective task, comparability between tasks, studies, and simulators can be achieved.

Keywords Minimally invasive surgery · Virtual reality trainer · Score · Skill assessment · Analytic hierarchy process · Delphi

The evolution of medical education in recent years has seen virtual reality (VR) simulation become a notable part of surgical education. Laparoscopic VR-trainers have gained attention [1] and are frequently used to provide a risk-free training environment [2, 3]. Most VR-trainers offer a set of basic skills tasks, giving trainees the opportunity to practice the most fundamental skills needed for laparoscopic surgery, such as hand–eye coordination, instrument coordination, and 2D–3D coordination. While VR-trainers have been shown to be equally as effective a training method as video trainers [4], they also carry a major advantage: the possibility for objective automated feedback. VR-trainers generate feedback automatically after each performance, eliminating the need for a supervisor to monitor a trainee's performance. This feedback consists of a variety of parameters; common ones include time, path length, error scores, and economy of movement [5, 6].

VR-trainer parameters have been subjected to many validation studies with equivocal results when treated as single parameters [7–10]. Therefore, one limitation of VR-trainer parameters is that they can be difficult to interpret when trying to gauge trainees' overall performance. For example,

✉ Felix Nickel
felix.nickel@med.uni-heidelberg.de

Extended author information available on the last page of the article

it is hard to know whether a performance with a shorter task time and a lower accuracy rate represents a higher skill level than one with a longer task time and a higher accuracy rate. A universally usable composite score incorporating the relevant parameters would address this problem and greatly improve the judgment of trainees' laparoscopic VR performance.

Some researchers and companies have already tried to incorporate VR-trainer parameters into different composite or cumulative scores. However, many of these scores include only two or three of all possible parameters [11, 12], sum parameters disregarding the different units [13, 14], or are based on and can only compare the specific groups used in a study [15]. Some manufacturers provide scores, but these are specific to their VR-trainer and validation and transparency of the score's computation are limited; the LapMentor™ score, for example, is based on unpublished expert performance data and the weighting of the different parameters is not accounted for [16]. Furthermore, the LapMentor™ score showed only partial construct validity for the first version of the LapMentor™ [17, 18] and to our knowledge has not yet been validated for the consecutive models.

A meaningful and standardized evaluation method to report VR data in research is therefore still lacking and greatly needed. Rosenthal et al. [19] highlight three important questions when analyzing VR data—"Which outcomes should be reported? How can outcomes be summarized and weighed? How can outcomes across different simulators and different studies be compared?".

This study addresses these questions and proposes a new, simulator-independent scoring standard—The Heidelberg VR Score. An international expert consensus on the importance of VR skill categories is established and a final score incorporating specific expert baseline data for both the LapMentor™ II and III (basic skill and peg transfer) is developed and validated.

Materials and methods

Setting and tasks

This study was conducted in the training center for Minimally Invasive Surgery of the Department of General, Visceral, and Transplantation Surgery at Heidelberg University Hospital, Germany. Participants received an introduction explaining type, extent, and value of this study before written informed consent was obtained. The study was approved by the local ethics committee at Heidelberg University (S-334/2011). The survey was performed using an online tool (http://www.umfrageonline.com).

A general scoring method for laparoscopic VR-trainer performance applicable to any simulator was established. This was applied to create final score for the laparoscopic VR-trainer LapMentor™ II and III (3D Systems, Rock Hill, USA), incorporating specific expert baseline data. LapMentor™ training scenarios used included Basic Skills tasks (eye–hand coordination, clipping, clipping and grasping, two-handed manoeuvres, cutting, and electrocautery) as well as the peg transfer task, which is part of the fundamentals of laparoscopic surgery (FLS) curriculum [20, 21]. The tasks were chosen to represent a wide variety of basic laparoscopic skills such as hand–eye coordination, bimanual dexterity, clip applying, cutting, and safe and accurate electrocautery.

For the development of a meaningful composite score, certain key steps had to be addressed:

(1) Identifying and categorizing relevant parameters
(2) Assigning weights to categories and parameters according to their importance
(3) Establishing a scoring algorithm
(4) Validating the final score.

Step 1 and 2: Identifying, categorizing, and weighting—integrated analytic hierarchy process-Delphi survey

To identify major skill categories and LapMentor™ parameters that reflect a good laparoscopic performance and to determine their weights within the composite score, a modified integrated analytic hierarchy process-Delphi survey was conducted. An outline of the proposed methodology can be found in Fig. 1.

Five main categories of general VR-trainer skills as well as relevant LapMentor™ parameters were identified based on the literature, reasoning of the authors and input from laparoscopically experienced surgeons. A Delphi-expert panel was then identified through published articles (inter alia with respect to research conducted with VR-trainers), laparoscopic expertise, and congress contributions. The expert panel was asked to compare the skill categories and LapMentor™ parameters pairwise using Saaty's 1–9 scale [22], resulting in judgment matrices. Individual judgment matrices were then checked for consistency and improved according to the algorithm proposed by Dong et al. [23], if no acceptable consistency was reached. Afterwards, a level of consensus between all experts was calculated. If a predefined level of consensus was not reached, another round of questionnaires was sent to the expert panel along with feedback on the results from the previous round. Experts were asked to reconsider their choices based on their fellow experts' opinions and comments. Once an acceptable level of consensus was reached or a significant improvement seen between two rounds, the Delphi survey was finished. An algorithm proposed by Dong et al. [23] was then used to reach a final consensus, and weights for each category and parameter were calculated. The names of the expert panel

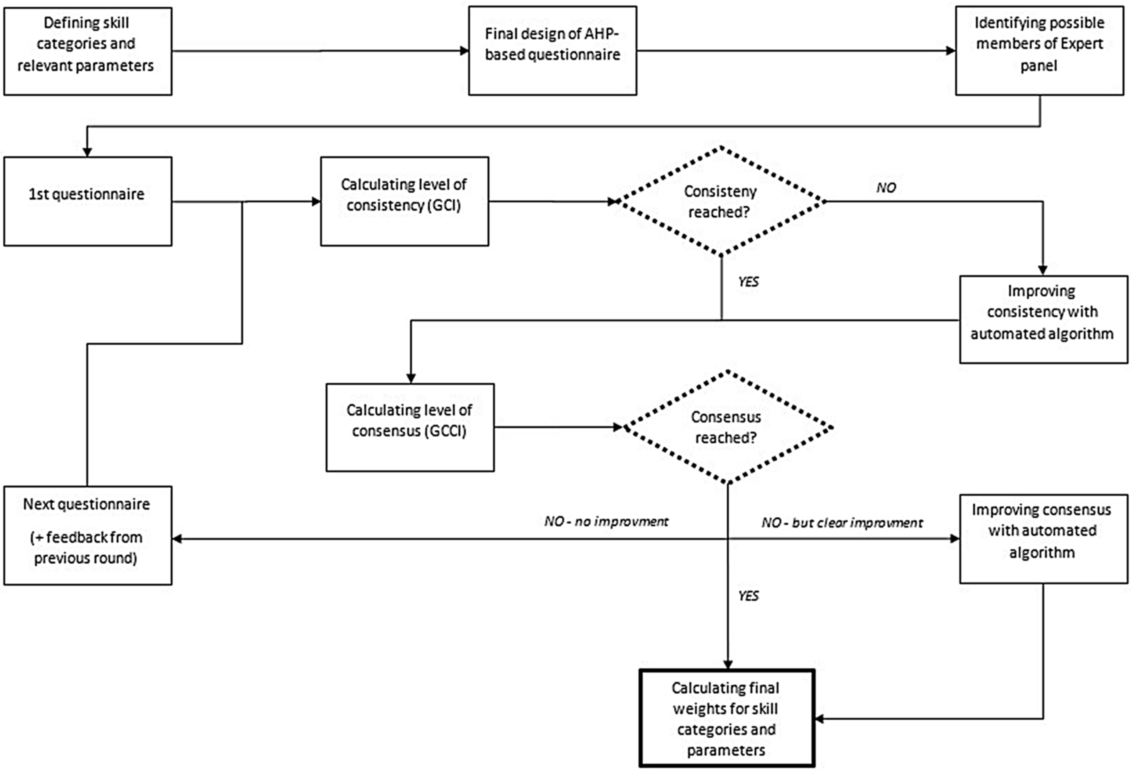
Defining skill categories and relevant parameters
Final design of AHP-based questionnaire
Identifying possible members of Expert panel
1st questionnaire
Calculating level of consistency (GCI)
Consisteny reached?
NO
YES
Improving consistency with automated algorithm
Calculating level of consensus (GCCI)
Consensus reached?
Next questionnaire
(+ feedback from previous round)
NO - no improvment
NO - but clear improvment
Improving consensus with automated algorithm
YES
Calculating final weights for skill categories and parameters

Fig. 1 A modified analytic hierarchy process—Delphi Survey. *AHP* analytic hierarchy process

remained undisclosed to all participants until a consensus was reached.

Step 3: Establishing a scoring method

To calculate the Heidelberg VR Score for a specific task, a set of expert baseline data for this task is needed. Four laparoscopic experts per LapMentor™ model were assessed to establish a baseline dataset for the seven tasks. After an initial warm-up period and an introduction to the tasks, all experts were asked to perform each task ten times. To support a high concentration and performance level, participants were allowed to take breaks between repetitions. A tutor was always present, to observe the performances and assist with any simulator-related difficulties if necessary.

Using the established expert dataset, expert mean $\left(\bar{M}_i^E\right)$ and trimmed standard deviation (σ_i^E) were calculated for each parameter (M_i). For the expert mean, only an expert's best performance for each parameter was used to set a high benchmark. To obtain the trimmed expert standard deviation, every value lower than two standard deviations from the original mean was excluded and the standard deviation recalculated using the remaining values. All parameters were then standardized using a z score statistic to account for the different scales and huge variability in standard deviations (e.g., time is measured in seconds with no upper limit, accuracy as a percentage with range limited between 0 and 100). The z score (Z_i) of a trainee's performance corresponds to their raw score (M_i^T) and expert data for the parameter (M_i) as follows:

$$Z_i = \frac{M_i^T - \bar{M}_i^E}{\sigma_i^E}.$$

The z scores of parameters were then adjusted to have a uniform direction, such that lower scores reflect a better performance (cost function). Parameters for which higher scores reflect a better performance needed to be reversed and were therefore multiplied by − 1.

Z scores and adjusted z scores for all relevant parameters were then aggregated into a standardized composite score for each task:

$$z = \left(\left(1 - \frac{\sum_{i=1}^{N} \alpha_i Z_i}{\sum_{i=1}^{N} \alpha_i Z_{\max}}\right) - Z_0\right) \times 100,$$

where N is the number of parameters, α_i is the weight associated with Z_i, andZ_0 is a measure of the outcome of the task based on Cotin et al. [24]. In this study, Z_{max} was set to 15 based on a data pool from over 90 laparoscopic novices performing all tasks three times. To increase readability, the standardized composite score was multiplied by 100.

Step 4: Validating the final, LapMentor™-specific composite score

Construct validity is often defined as whether or not a tool measures the trait it is supposed to measure [25]. In training research, it is commonly assessed through a tool's ability to differentiate between different levels of laparoscopic experience [26]. To validate the composite score, laparoscopic experts and novices were assessed on the LapMentor™ II and III, performing all seven tasks twice. Participants were categorized according to the number of laparoscopic procedures performed (Experts > 100; Novices = 0). None of the experts used for validation were included in the initial baseline dataset.

Statistical analysis and sample size calculation

Individual judgment matrices were checked for consistency using the geometric consistency index (GCI). A GCI of ≤ 0.35 was regarded as acceptable consistency [27]. The level of consensus was calculated using the geometric cardinal consensus index (GCCI) [23]. Based on the work of Dong et al. [23] and Aguarón et al. [28], a GCCI of ≤ 0.35 was considered an acceptable level of consensus. Improvement of the GCCI during the Delphi survey was tested using the Wilcoxon Signed-rank test. A p value of less than 0.05 was considered statistically significant. Final weights were derived from the judgment matrices using the row geometric mean prioritization method (RGMM) [27] under the aggregation of individual judgments (AIJ). Pairwise comparisons using the Saaty Scale were reported using geometric mean and geometric standard deviation. Arithmetic mean and standard deviation were calculated for the deduction per failed outcome measure. To establish construct validity, composite scores were compared between experts and novices using a Mann–Whitney U test. Statistical analysis was performed using SPSS version 24.0 (IBM SPSS Statistics, IBM Corp.). GCI, GCCI, and improvement algorithms were implemented in a Java script.

For the validation study, a conservative sample size calculation was conducted with respect to a preliminary composite score of the peg transfer task. The calculation was based on performances from a previous study in a similar setting (Experts: 98.54 ± 9.20; Novices: 65.89 ± 10.96). With a two-sided significance level $\alpha = 0.05$ and a power of $1 - \beta = 0.8$, a conservative sample size calculation resulted in nine participants per group. Due to the small sample sizes required and the great inhomogeneity usually seen in novice performances, we planned to exceed this number by six in the Novice group.

Results

AHP-integrated Delphi survey

Five major and simulator-independent categories were identified (Table 1). For the LapMentor™-specific score established in this paper, a final set of relevant LapMentor™ parameters was chosen and each allocated to a category (Table 2). A total of 14 eligible experts (a list of members can be found in the acknowledgements) accepted the invitation to participate in the Delphi survey, all of whom went on to participate in the second survey round. The GCCI decreased significantly ($p = 0.002$) from the first to the second round, suggesting an increased level of consensus and the Delphi survey was finished. More detailed results of the Delphi survey can be found in Fig. 2 and Tables 1 and 2.

Four of the five categories (time, safety, efficiency, dexterity) were assigned a weight. For the fifth category, outcome parameters, the expert panel was asked to define a specific value (possible range 0–100) or ratio to be deducted from the final score (Z_0). All experts agreed to deduct the quotient of failed outcome items divided by possible outcome items, e.g., if the task requires nine ducts to be clipped for its completion and only seven are clipped, a value of 2/9 (= 0.2222) will be deducted from the final score. If two or more parameters were allocated to the same category for any given task, the expert panel was asked to compare them pairwise to determine their relative importance and assign them weights within the category (Table 1). If a task had no parameters assigned to a certain category, the weights of the remaining skill categories were newly calculated (see Table 3 in Appendix).

Figure 2 shows the development of weights assigned to each category during the survey process. While "Safety" increased in importance, all other categories decreased in importance from the first to the second round. Final weights, derived after application of the consensus improving algorithm, differ only slightly from the original expert judgments in round two.

Table 1 Definition and weights of main skill categories

Category	Definition	Weighting (%)
Time	Task time	5.7
Efficiency	Unnecessary actions	17.3
Safety	Complications/damage	67.3
Dexterity	Manual skill	9.7
Outcome	Errors preventing achievement of the task's goal	Failed/total possible items[a]

[a] For the category outcome a fixed value (Z_0) was deducted from the final score

Table 2 LapMentor™ tasks and categorized parameters chosen by the author panel and laparoscopic experts during the Delphi process; specific weights for LapMentor™ parameters established through the analytic hierarchy process-Delphi survey

Task	Categories with included parameters	Specific weight inside category[b]
General (all tasks)	Time	
	Total time	100
	Dexterity	
	Total path length[a]	37.2
	Total number of movements[a]	62.8
Eye–hand coordination	Safety	
	Accuracy rate—touched targets (%)	100
	Task outcome	
	Number of correct hits	100
Clip applying/clipping and grasping	Efficiency	
	Accuracy rate—applied clips (%)	100
	Task outcome	
	Number of clipped ducts	100
Two-handed maneuvers	Task outcome	
	Number of exposed green balls that are collected	100
Cutting	Safety	
	Number of cutting maneuvers performed that cause injuries[a]	68.6
	Number of retraction operations with overstretch injuries to the tissue[a]	31.4
	Efficiency	
	Total number of cutting maneuvers	47.3
	Total number of retraction operations	52.7
Electrocautery	Safety	
	Time cautery is applied on non-highlighted bands	100
	Efficiency	100
	Efficiency of cautery (%)	
	Task outcome	
	Number of highlighted bands that were cut	100
Peg transfer	Task outcome	
	Pegs transferred	100

[a] Newly calculated parameter from two existing parameters

[b] For categories containing more than one parameter, the expert panel was asked to compare the importance of each parameter in this category. Specific weights were then calculated for each parameter based on the expert judgment

Validation study

Participants totaled to 24 LapMentor™ II and 23 on the LapMentor™ III, with Expert groups of 9 per simulator. Participants completed all seven tasks twice. For the second attempt, Experts outperformed Novices significantly in all tasks on both the LapMentor™ II and III, except for the task Electrocautery on the LapMentor™ III (Figs. 3, 4). The first attempt showed comparable results; the score was able to differentiate between Experts and Novices for all tasks on both VR-trainers, except Electrocautery on the LapMentor™ III.

Discussion

This study presents the creation of a universally applicable composite score for VR training in laparoscopic surgery, the Heidelberg VR score. The score's formulation was guided by the consensus of an international expert panel. Furthermore, a complete score was created and validated for basic skills tasks and peg transfer on the LapMentor™ II and III. A website (heidelbergvrscore.de) is being created to allow other researchers and trainees to easily calculate the Heidelberg VR score for their own data and simulators, given the availability of expert data on the parameters and tasks.

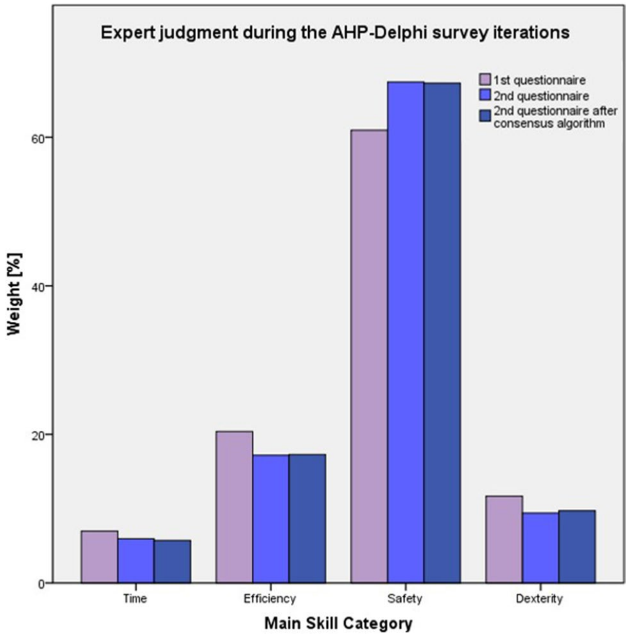

Fig. 2 Development of weights for main skill categories during integrated analytic hierarchy process (AHP)-Delphi survey

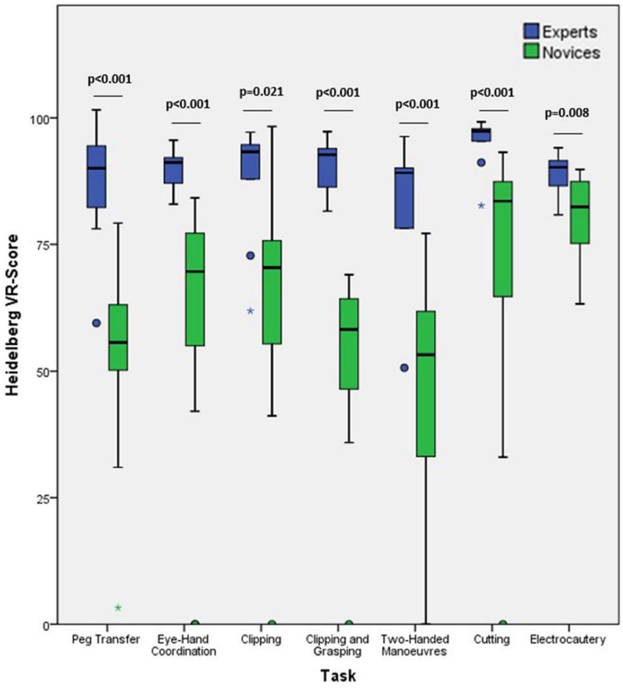

Fig. 3 Expert and Novice Heidelberg VR scores for single tasks on the LapMentor™ II, 2nd attempt

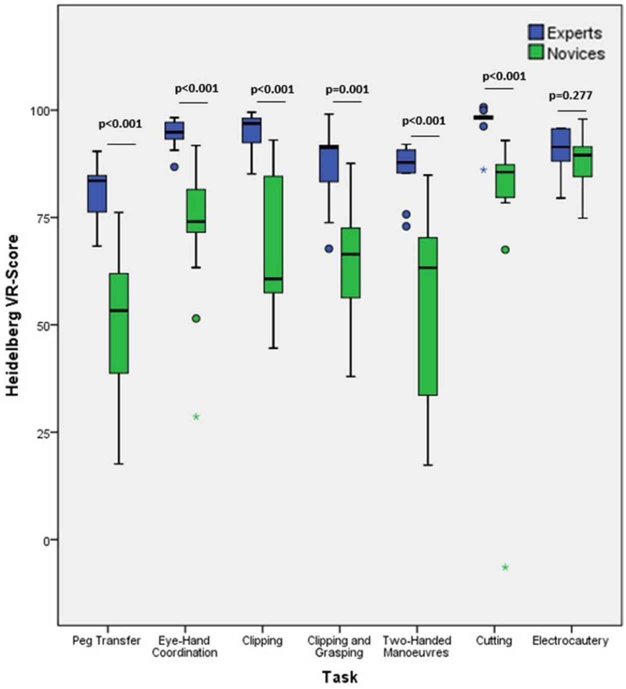

Fig. 4 Expert and Novice Heidelberg VR scores for single tasks on the LapMentor™ III, 2nd attempt

A combination of two well-established methodologies was used to reach expert consensus on the importance of different skill aspects in laparoscopic surgery. The first was the Delphi method. This method, as proposed by Dalkey et al., is an iterative survey process based on anonymity, controlled feedback, and statistical group response [29] and has been used in many areas of research, including in the field of medicine [30–32]. The second method used was the AHP, originally introduced by Saaty [22], and works with the assumption that complex ideas can be judged more effectively through establishment of a hierarchy. In this study, this was done by identifying and weighting five major skill categories that contribute to a good laparoscopic performance. Rather than considering all available parameters for one task individually, they were categorized and looked at as part of a hierarchy. This hierarchical structure is of great value as weighting the different parameters of different tasks allows for transferability of the score. Once a parameter is allocated to a category, it is automatically assigned a weight (that already given to the category), removing the need for another expert consensus. Furthermore, the AHP relies on pairwise comparisons, which are considered to reflect one's opinion more precisely than comparing multiple items at a time [33].

The Delphi-expert panel in this study consisted of fourteen members. While there is no consensus in the literature on the number of participants needed for such a panel, this number lies within the upper range of the 8–15 suggested in many papers for homogeneous groups with similar qualifications [34, 35]. No drop outs were recorded, which is rare in the literature [36]. The panel went through two actual Delphi iterations, with the questionnaire designed beforehand. This is in accordance with the suggested iterations for a Delphi survey in health research [37]. To eliminate the time and resource constraints of conducting multiple Delphi iterations, the consensus achieving algorithm proposed by Dong et al. [23] was applied. This slowly adjusts the individual expert judgments, moving them towards consensus. The algorithm was applied after a significant increase in the level of consensus could be seen between Delphi iterations, indicating that the expert opinions had converged. Although the algorithm increased the level of consensus to an acceptable level, the final weights allocated to each category only varied slightly following its application (see Fig. 2). This proves that individual expert judgment is the major contributor to the final consensus, while the algorithm only helps to reach it.

The five major skill categories were chosen with the assumption that any parameter from any task could be sorted into one of them. The following were criteria for parameter exclusion: (1) it cannot differentiate a good from a bad performance, (2) it is already represented by other parameters in the score (to avoid overrepresentation of a certain aspect), and (3) its calculation is unprecise or unclear. The manufacturer was contacted to discuss unclear calculations, but if no sufficient explanation was given, the parameter was excluded. Two new parameters were calculated based on existing parameters (see Table 2). This was done to turn the original parameters into unidirectional functions, with lower values representing a better performance.

The final results of the Delphi questionnaire indicate that the skill category “Safety” is of tremendous importance to the Experts, whereas the time needed to complete a task is of least importance. This is in accordance with the results of an expert consensus on robotic surgical parameters [38], where safety and critical errors were rated the most important. To our knowledge, ours is the first study to establish an expert consensus on the relative importance within major skill categories of parameters contributing to a good laparoscopic performance. In 2004, Stylopoulos et al. administered a survey asking 30 laparoscopic surgeons to rate various skill parameters for their computer-enhanced laparoscopic training system [39]. However, since they focused on rating individual parameters, their results cannot be transferred to other VR-trainers reporting different parameters. The categorization of parameters into major skill categories presented in this paper allows researchers to work with and calculate weighted scores on any VR-trainer, provided an expert baseline dataset is available.

The Heidelberg VR Score offers the highly needed opportunity to compare trainees not only on one task but across different tasks on a simulator, between studies, and even between different simulators. As mentioned above, the score can be transferred to any VR-trainer task given an expert baseline dataset. This expert data is used to standardize task performance, making comparisons between tasks possible; any specific score, for example 80, will always represent the same distance from the expert baseline and can be taken as equivalent across tasks. This is important in addressing the different foci of tasks; the LapMentor™ Peg transfer task, for example, requires a relatively long path length for task completion (all pegs transferred) compared to others. Therefore, to transfer the Heidelberg VR score to any simulator or task only an expert baseline dataset is needed for the specific task on the specific simulator. To ensure comparability, a large expert baseline dataset is favorable.

The calculation of the Heidelberg VR score is based on the work of Cotin et al., who presented a universal scoring system for computer-assisted laparoscopic skills training [24] and Rosenthal et al., who proposed a general evaluation method [19]. Cotin et al.’s 2002 score used a standardized set of task-independent skill assessment parameters, which could technically be implemented into any VR-trainer [24]. However, widely distributed VR-trainers such as the LapMentor™ report task-specific outcomes and do not report all parameters proposed and used by Cotin et al., such as motion smoothness or depth perception. This might be a reason why their composite score has not achieved general usage. Rosenthal et al. recently proposed a new standard for reporting outcome measures for VR-trainers meant to address the risk of selective outcome reporting and multiplicity issues [19]. While they proposed an outline applicable to any simulator, their research did not focus on determining weights for specific laparoscopic skills. Therefore, their scoring method cannot be applied without additional work, such as that presented in this study.

To highlight the importance of outcome parameters, defined as critical errors that hinder a trainee from accomplishing the goal of a task, a specific value is deducted from the final Heidelberg VR score. If transferred to a clinical setting, these errors would include major complications that worsen the outcome of an operation and cannot be reversed or controlled. Initially, the expert panel was asked to deduct a fixed value; the consensus suggested subtracting 10 points on the scoring scale (where 100 is the mean of the best expert performance scores) for every failed item. However, it was suggested by a few panel members to deduct a ratio of failed items divided by possible items for each task. This was suggested in the second Delphi round and preferred by 100% of the members of the expert panel. Therefore, the scoring method was adjusted accordingly.

The final Heidelberg VR score for the LapMentor™ II and III was able to differentiate between laparoscopic Novices and Experts, demonstrating construct validity. However, the mean difference between Expert and Novice scores varies between the tasks. This suggests that some basic skills tasks such as “Electrocautery” may have been too easy, leading to no trainee mistakes and preventing effective assessment of small variations in their skill levels. A study by Agrusa et al. comparing 3D and 2D laparoscopic surgery showed a similar phenomenon; significant differences were seen only for complex surgical procedures [40]. No significant differences between experts and novices could be found for the task "Electrocautery" on the LapMentor™ III, though they could be for the same task on the LapMentor™ II. When analyzing the results in more detail, it could be seen that this was due to an unsafe working technique used by three of the experts on the LapMentor™ III; an extraordinarily high amount of time was spent cauterizing non-highlighted bands compared to their peers. Safety is weighted highly in the composition of the Heidelberg VR score, leading this mistake to reduce these experts' performance scores on the task. One explanation for this unsafe operation technique could be that surgeons, who usually operate on human patients, might not value the aspect of safety on a simulator as much and prefer to operate as efficiently as possible, while taking more risks than usual. We expect the Heidelberg VR score to assess trainee performance on this task correctly given instructions are followed by experts on how best to perform VR tasks.

Limitations

This study has some limitations to be considered when interpreting the results. Some studies have highlighted methodological weaknesses of the Delphi methodology. These include, for example, the subjectivity of defining who is an expert [37], the limited possibilities for expert interaction due to the feedback being written and controlled [41], as well as the time needed to carry out the whole process. Nevertheless, experts can be selected carefully, and the benefits of the Delphi methodology continue to outweigh the risks for this study. The anonymity it requires supports the use of modern communication tools instead of in-person meetings to achieve an expert consensus, simplifying the process. Furthermore, it prevents the opinion of more dominant characters from gaining more attention than those of a shy character [42]. Additionally, sending back group feedback informs all panel members of the current state of opinion and highlights objectives possibly otherwise unconsidered by some members. Therefore, the Delphi methodology explicitly offers the opportunity to change one's mind [43]. It should be noted that the Heidelberg VR Score does not improve or alter the measurements of parameters offered by a VR-trainer, but rather creates a uniform evaluation based on them and expert baseline data. This, however, offers the highly needed comparability between different simulators and tasks.

Conclusion

In conclusion, a standardized reporting and evaluation system for laparoscopic VR performances is of utmost importance to allow comparability, increase the relevance and interpretation of reported outcomes, and minimize the risk of selective outcome reporting. Such a new standard is proposed with the Heidelberg VR Score, which addresses the shortcomings of the wide variety of reporting methods available. A final, practically useable score for the basic skills and peg transfer tasks on the LapMentor™ II and III was established and validated. To facilitate use of the Heidelberg VR Score for other researchers and training centers, a website (heidelbergvrscore.de) is being created to simplify score calculation. To establish the Heidelberg VR score as a new standard for all VR-trainers, expert baseline data for new simulators and tasks must be collected in future research for integration into and evaluation of the score.

Acknowledgements The authors would like to thank all members of the expert panel for their support: Esther Bonrath, Germany; Sanne Botden, Netherlands; Julian Bucher, Germany; Dieter Hahnloser, Switzerland, Daniel A. Hashimoto, USA; Tobias Huber, Germany; Georg Linke, Switzerland; Sören Torge Mees, Germany; Daniel Miscovic, UK; Christoph Reißfelder, Germany; Marlies Schijven, Netherlands; Lee Swanström, France; Siska van Bruwane, Belgium; Markus Wallwiener, Germany. Furthermore, we would like to thank Hubertus Feußner, Laurents Stassen, and Thomas Vogel for sharing their experience for this project. Furthermore, the authors would like to thank Mr. Nicolas Billen for his help with implementing the scoring algorithm, Mr. Samuel Kilian for his help during the calculation process, and Ms. Linhong Li for her help with setting up the website.

Compliance with ethical standards

Disclosure Mona W. Schmidt, Karl-Friedrich Kowalewski, Marc L. Schmidt, Erica Wennberg, Carly R. Garrow, Sang Paik, Laura Benner, Marlies Schijven, Beat-Peter Müller Stich, and Felix Nickel have no conflict of interest or financial ties to disclose.

Appendix

Because basic skill tasks on VR-trainers often want to test a specific skill (e.g., eye–hand coordination), some tasks do not include any relevant parameters for some of the main skill categories (e.g., no relevant safety parameter in the task clip applying). Therefore, a new calculation of the importance of the remaining skill categories compared to each other is necessary. Due to the nature of the analytic hierarchy process, with its pairwise comparison, this does not require a new expert judgment. The weights can easily be calculated using only the pairwise comparisons of the

Table 3 Weights (%) of main skill categories for all possible combinations of main skill categories in a task

Missing categories	Time	Efficiency	Safety	Dexterity
Time	No parameter assigned to this category	15.7	75.9	8.4
Efficiency	8.5	No parameter assigned to this category	76.8	14.7
Safety	13.4	59.0	No parameter assigned to this category	27.6
Dexterity	6.9	17.1	76.0	No parameter assigned to this category
Time + efficiency	No parameter assigned to this category	No parameter assigned to this category	87.1	12.9
Time + safety	No parameter assigned to this category	73.6	No parameter assigned to this category	26.4
Time + dexterity	No parameter assigned to this category	13.2	86.8	No parameter assigned to this category
Efficiency + safety	27.0	No parameter assigned to this category	No parameter assigned to this category	73.0
Efficiency + dexterity	12.2	No parameter assigned to this category	87.8	No parameter assigned to this category
Safety + dexterity	19.7	80.3	No parameter assigned to this category	No parameter assigned to this category

All weights reported in percent

remaining skill categories for the task. Therefore, we used the expert judgments from the second Delphi iteration of the remaining skill categories and applied consistency and consensus improving algorithms as described above for each possible combination of skill categories. The weights of all possible combinations of categories can be found in Table 3 in Appendix.

References

1. Buckley CE, Nugent E, Ryan D, Neary PC (2012) Virtual reality—a new era in surgical training. In: Eichenberg C (ed) Virtual reality in psychological, medical and pedagogical applications. InTech, Rijeka. https://doi.org/10.5772/46415
2. Yiannakopoulou E, Nikiteas N, Perrea D, Tsigris C (2015) Virtual reality simulators and training in laparoscopic surgery. Int J Surg 13:60–64
3. Nickel F, Brzoska JA, Gondan M, Rangnick HM, Chu J, Kenngott HG, Linke GR, Kadmon M, Fischer L, Muller-Stich BP (2015) Virtual reality training versus blended learning of laparoscopic cholecystectomy: a randomized controlled trial with laparoscopic novices. Medicine 94:e764
4. Alaker M, Wynn GR, Arulampalam T (2016) Virtual reality training in laparoscopic surgery: a systematic review & meta-analysis. Int J Surg 29:85–94
5. Kowalewski KF, Garrow CR, Proctor T, Preukschas AA, Friedrich M, Muller PC, Kenngott HG, Fischer L, Muller-Stich BP, Nickel F (2018) LapTrain: multi-modality training curriculum for laparoscopic cholecystectomy-results of a randomized controlled trial. Surg Endosc 32:3830–3838
6. Beyer-Berjot L, Berdah S, Hashimoto DA, Darzi A, Aggarwal R (2016) A virtual reality training curriculum for laparoscopic colorectal surgery. J Surg Educ 73:932–941
7. Thijssen AS, Schijven MP (2010) Contemporary virtual reality laparoscopy simulators: quicksand or solid grounds for assessing surgical trainees? Am J Surg 199:529–541
8. Yamaguchi S, Konishi K, Yasunaga T, Yoshida D, Kinjo N, Kobayashi K, Ieiri S, Okazaki K, Nakashima H, Tanoue K, Maehara Y, Hashizume M (2007) Construct validity for eye-hand coordination skill on a virtual reality laparoscopic surgical simulator. Surg Endosc 21:2253–2257
9. Aggarwal R, Crochet P, Dias A, Misra A, Ziprin P, Darzi A (2009) Development of a virtual reality training curriculum for laparoscopic cholecystectomy. Br J Surg 96:1086–1093
10. Wilson M, McGrath J, Vine S, Brewer J, Defriend D, Masters R (2010) Psychomotor control in a virtual laparoscopic surgery training environment: gaze control parameters differentiate novices from experts. Surg Endosc 24:2458–2464
11. Schijven M, Jakimowicz J (2003) Construct validity: experts and novices performing on the Xitact LS500 laparoscopy simulator. Surg Technol Int 11:32–36
12. Larsen CR, Grantcharov T, Aggarwal R, Tully A, Sorensen JL, Dalsgaard T, Ottesen B (2006) Objective assessment of gynecologic laparoscopic skills using the LapSimGyn virtual reality simulator. 20:1460–1466
13. Van Sickle KR, Ritter EM, McClusky DA III, Lederman A, Baghai M, Gallagher AG, Smith CD (2007) Attempted establishment of proficiency levels for laparoscopic performance on a national scale using simulation: the results from the 2004 SAGES minimally invasive surgical trainer-virtual reality (MIST-VR) learning center study. Surg Endosc 21:5–10
14. Avgerinos DV, Goodell KH, Waxberg S, Cao CG, Schwaitzberg SD (2005) Comparison of the sensitivity of physical and virtual

laparoscopic surgical training simulators to the user's level of experience. Surg Endosc 19:1211–1215

15. van Dongen KW, Tournoij E, van der Zee DC, Schijven MP, Broeders IA (2007) Construct validity of the LapSim: can the LapSim virtual reality simulator distinguish between novices and experts? Surg Endosc 21:1413–1417
16. Andreatta PB, Woodrum DT, Gauger PG, Minter RM (2008) LapMentor metrics possess limited construct validity. Simul Healthc 3:16–25
17. Zhang A, Hunerbein M, Dai Y, Schlag PM, Beller S (2008) Construct validity testing of a laparoscopic surgery simulator (Lap Mentor): evaluation of surgical skill with a virtual laparoscopic training simulator. Surg Endosc 22:1440–1444
18. McDougall EM, Corica FA, Boker JR, Sala LG, Stoliar G, Borin JF, Chu FT, Clayman RV (2006) Construct validity testing of a laparoscopic surgical simulator. J Am Coll Surg 202:779–787
19. Rosenthal R, von Websky MW, Hoffmann H, Vitz M, Hahnloser D, Bucher HC, Schäfer J (2015) How to report multiple outcome metrics in virtual reality simulation. Eur Surg 47:202–205
20. Okrainec A, Soper NJ, Swanstrom LL, Fried GM (2011) Trends and results of the first 5 years of fundamentals of laparoscopic surgery (FLS) certification testing. Surg Endosc 25:1192–1198
21. Peters JH, Fried GM, Swanstrom LL, Soper NJ, Sillin LF, Schirmer B, Hoffman K, the SFLSC (2004) Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery 135:21–27
22. Saaty TL (1980) The analytic heirarchy process: planning, priority setting, resource allocation. McGraw-Hill, New York
23. Dong Y, Zhang G, Hong W, Xu Y (2010) Consensus models for AHP group decision making under row geometric mean prioritization method. Decis Support Syst 49:281–289
24. Cotin S, Stylopoulos N, Ottensmeyer M, Neumann P, Rattner D, Dawson S (2002) Metrics for laparoscopic skills trainers: the weakest link! In: Dohi T, Kikinis R (eds) Medical image computing and computer-assisted intervenetion—MICCAI. Springer, Berlin, Heidelberg, pp 35–43
25. Moorthy K, Munz Y, Sarker SK, Darzi A (2003) Objective assessment of technical skills in surgery. BMJ 327:1032–1037
26. McDougall EM (2007) Validation of surgical simulators. J Endourol 21:244–247
27. Crawford G, Williams C (1985) A note on the analysis of subjective judgment matrices. J Math Psychol 29:387–405
28. Aguaron J, Moreno-Jiménez JMa (2003) The geometric consistency index: approximated thresholds. Eur J Oper Res 147:137–145
29. Dalkey NC (1969) The Delphi method: an experimental study of group opinion RAND CORP SANTA MONICA CALIF
30. Landeta J (2006) Current validity of the Delphi method in social sciences. Technol Forecast Soc Chang 73:467–482
31. Awad M, Awad F, Carter F, Jervis B, Buzink S, Foster J, Jakimowicz J, Francis NK (2018) Consensus views on the optimum training curriculum for advanced minimally invasive surgery: a delphi study. Int J Surg 53:137–142
32. Cuschieri A, Francis N, Crosby J, Hanna GB (2001) What do master surgeons think of surgical competence and revalidation? 1. Am J Surg 182:110–116
33. Blumenthal AL (1977) The process of cognition. Experimental psychology series. Prentice Hall/Pearson Education, New Jersey
34. Palter VN, Graafland M, Schijven MP, Grantcharov TP (2012) Designing a proficiency-based, content validated virtual reality curriculum for laparoscopic colorectal surgery: a Delphi approach. Surgery 151:391–397
35. Skulmoski GJ, Hartman FT, Krahn J (2007) The Delphi method for graduate research. J Inf Technol Educ 6:1–21
36. Keeney S, McKenna H, Hasson F (2010) The Delphi technique in nursing and health research. Wiley, Chichester
37. Trevelyan EG, Robinson PN (2015) Delphi methodology in health research: how to do it? Eur J Integr Med 7:423–428
38. Chowriappa AJ, Shi Y, Raza SJ, Ahmed K, Stegemann A, Wilding G, Kaouk J, Peabody JO, Menon M, Hassett JM, Kesavadas T, Guru KA (2013) Development and validation of a composite scoring system for robot-assisted surgical training—the Robotic skills assessment score. J Surg Res 185:561–569
39. Stylopoulos N, Cotin S, Maithel SK, Ottensmeyer M, Jackson PG, Bardsley RS, Neumann PF, Rattner DW, Dawson SL (2004) Computer-enhanced laparoscopic training system (CELTS): bridging the gap. Surg Endosc 18:782–789
40. Agrusa A, Di Buono G, Buscemi S, Cucinella G, Romano G, Gulotta G (2018) 3D laparoscopic surgery: a prospective clinical trial. Oncotarget 9:17325
41. Milkovich G, Annoni AJ, Mahoney T (1972) The use of the Delphi procedures in manpower forecasting. Manag Sci 19:381–388
42. Khorramshahgol R, Moustakis V (1988) Delphic hierarchy process (DHP): a methodology for priority setting derived from the Delphi method and analytical hierarchy process. Eur J Oper Res 37:347–354
43. Hsu C-C, Sandford BA (2007) The Delphi technique: making sense of consensus. Pract Assess Res Eval 12:1–8

Affiliations

Mona W. Schmidt[1] · Karl-Friedrich Kowalewski[1] · Marc L. Schmidt[2] · Erica Wennberg[1] · Carly R. Garrow[1] · Sang Paik[1] · Laura Benner[3] · Marlies P. Schijven[4] · Beat P. Müller-Stich[1] · Felix Nickel[1]

Mona W. Schmidt
mona.schmidt@med.uni-heidelberg.de

Karl-Friedrich Kowalewski
karl-friedrich.kowalewski@med.uni-heidelberg.de

Erica Wennberg
erica.wennberg@outlook.com

Carly R. Garrow
crggx6@mail.missouri.edu2

Sang Paik
sang.paik@med.uni-heidelberg.de

Laura Benner
benner@imbi.uni-heidelberg.de

Marlies P. Schijven
m.p.schijven@amc.uva.nl

Beat P. Müller-Stich
beat.mueller@med.uni-heidelberg.de

[1] Department of General, Visceral, and Transplantation Surgery, Heidelberg University Hospital, Im Neuenheimer Feld 110, 69120 Heidelberg, Germany

[2] Karlsruhe, Germany

[3] Department of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany

[4] Deparment of Surgery, Amsterdam Gastroenterology and Metabolism, Amsterdam UMC, University of Amsterdam, PO Box 22660, 1100 DD Amsterdam, The Netherlands