Improving outcomes in pediatric endoscopic third ventriculostomy through outcome analysis and surgeon training
Breimer, Gerben Eise

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 01-11-2019
Chapter 5

Validity Evidence for the Neuro-Endoscopic Ventriculostomy Assessment Tool (NEVAT)

Gerben E. Breimer
Faizal A. Haji
Giuseppe Cinalli
Eelco W. Hoving
James M. Drake

Operative Neurosurgery 2016
DOI: 10.1227/NEU.0000000000001158
Abstract

Background
Growing demand for transparent and standardized methods for evaluating surgical competence prompted the construction of the Neuro-Endoscopic Ventriculostomy Assessment Tool (NEVAT).

Objective
To provide validity evidence of the NEVAT by reporting on the tool’s internal structure and its relationship with surgical expertise during simulation-based training.

Methods
The NEVAT was used to assess performance of trainees and faculty at an international neuroendoscopy workshop. All participants performed an endoscopic third ventriculostomy (ETV) on a synthetic simulator. Participants were simultaneously scored by 2 raters using the NEVAT procedural checklist and global rating scale (GRS). Evidence of internal structure was collected by calculating interrater reliability and internal consistency of raters’ scores. Evidence of relationships with other variables was collected by comparing the ETV performance of experts, experienced trainees, and novices using Jonckheere’s test (evidence of construct validity).

Results
Thirteen experts, 11 experienced trainees, and 10 novices participated. The interrater reliability by the intraclass correlation coefficient for the checklist and GRS was 0.82 and 0.94, respectively. Internal consistency (Cronbach’s a) for the checklist and the GRS was 0.74 and 0.97, respectively. Median scores with interquartile range on the checklist and GRS for novices, experienced trainees, and experts were 0.69 (0.58-0.86), 0.85 (0.63-0.89), and 0.85 (0.81-0.91) and 3.1 (2.5-3.8), 3.7 (2.2-4.3) and 4.6 (4.4-4.9), respectively. Jonckheere’s test showed that the median checklist and GRS score increased with performer expertise (P = .04 and .002, respectively).
Conclusion
This study provides validity evidence for the NEVAT to support its use as a standardized method of evaluating neuroendoscopic competence during simulation-based training.

Introduction
Significant changes in the landscape of surgical education, including resident duty hour restrictions, concerns about patient safety, mounting costs of health care, and the shift toward competency-based education, are putting significant strain on traditional methods of training and assessing residents’ surgical skills. These factors, along with the increase in competency-based education, have led to growing interest in the use of simulation in postgraduate surgical education.\textsuperscript{1-5} Whether simulation is used for training or to evaluate the skills of trainees at various points during residency, the need for assessment tools that are valid, reliable, and easily implemented has never been greater.

Contemporary approaches to the development and evaluation of assessment tools in health professions education view validity as an argument, whereby evidence is collected to support the use of an assessment instrument in a particular setting.\textsuperscript{6} A widely cited framework describes 5 different types of evidence that can be used to support the validity of scores generated from an assessment tool: content, response process, internal structure, relationships with other variables, and consequences.\textsuperscript{7} Depending on the intended use of the instrument (eg, formative assessment vs credentialing decision), the level of evidence needed to support the tool will change. As such, the process of validation is not considered to be a one-time event.\textsuperscript{6,8,9} For example, if we demonstrate that 1 or more assessment scales are “valid” for assessing junior resident performance during simulation-based training of a given procedure, this does not necessarily translate into evidence supporting the use of these tools in the assessment of senior residents in the operating room (OR).\textsuperscript{10}

In a previous study by our research group, a series of assessment instruments for the endoscopic third ventriculostomy (ETV) procedure were constructed to fulfill the need for standardized assessment methods for ventricular neuroendoscopy.\textsuperscript{11} Together these are called the Neuro-Endoscopic Ventriculostomy Assessment Tool.
(NEVAT). We previously generated content validity evidence for the NEVAT through a Delphi study of an international panel of expert neuroendoscopists. However, before these tools can be readily implemented in the assessment of ETV performance, further validity evidence related to internal structure and to relationships with other variables is required. The phrase “relationship to other variables” is a form of validity evidence in the framework that we used. It measures whether scores from the assessment instrument generate similar results as would be predicted from other measures of the underlying construct. In this study, it refers to whether the NEVAT scores can differentiate between the expertise of performers that would be predicted based on their previous experience with ETV (also known as construct validity).

The purpose of this study is twofold. First, we sought to provide validity evidence related to internal structure by examining the interrater reliability and internal consistency of scores generated using the NEVAT assessment tools. Second, we sought to examine NEVAT’s relationships with other variables by comparing scores observed among novices, experienced trainees, and expert surgeons performing ETV on a physical simulator. Both of these aims were realized by collecting data on the NEVAT during an international neuroendoscopy workshop.

Methods

Workshop and Participants

The study took place during the annual international hands-on workshop on cerebral and ventricular neuroendoscopy at the Centre of Biotechnologies, Cardarelli Hospital, Naples, Italy (January 2014). All faculty were expert neuroendoscopists. The participants were a mix of neurosurgical residents, fellows, and consultant neurosurgeons. Preceding the course, participants and faculty alike received an invitation letter with information on the goals and format of the study.

Thirty-four participants were included in the study, 13 of whom were faculty and 21 of whom were trainees. At the outset of the workshop, participants completed a demographic survey in which data regarding their status (resident [plus postgraduate year], fellow, or consulting neurosurgeons), previous experience...
Validity evidence for NEVAT

with ETV (observed, performed with assistance, and performed independently), and previous exposure to simulation-based training. The participants were subsequently divided into 3 groups: novices, “experienced trainees,” and experts. Novices were defined as assisting in fewer than 5 ETVs and performing none independently. The “experienced trainees” were defined as assisting in 5 or more ETVs and/or performing at least 1 independently. All faculty were regarded as experts.

**Study Protocol**

After providing informed consent, all participants were given the opportunity to perform an ETV procedure on a synthetic phantom, the S.I.M.O.N.T. Neurosurgical Endotrainer (Sinus Model Oto-Rhino Neuro Trainer, Pro Delphus Co.). Participants were instructed to complete a full ETV procedure as they would in the OR; all steps that could not be carried out on the simulator had to be mentioned as if they also were physically executed. Standard neuroendoscopic instruments were used during the simulation, including a Gaab universal rigid 0° endoscope (Karl-Storz, Tuttlingen, Germany), trocar, and grasping and dissecting forceps. Irrigation with normal saline solution was also provided. Participants were videotaped while performing the ETV in an anonymous manner (ie, by only videotaping hands and screen plus operating site, without including the face of the participant in the video).

**Assessment Tool and Rating Procedure**

The NEVAT contains 3 assessment measures: a procedure-specific checklist, a checklist of surgical errors, and a global rating scale (GRS).\(^{11}\) Checklist items were graded on a binary scale (done correctly vs done incorrectly) and the GRS items were graded on a 5-point Likert scale (with explicit verbal anchors associated with ratings of 1, 3, and 5). The GRS and procedure-specific checklist of the NEVAT were used for scoring all participants consisting of, respectively, 42 and 10 items (see Appendix, Supplemental Digital Content, http://links.lww.com/NEU/A820). A number of checklist items could not be completed on the simulator (corrected checklist with nonscorable items removed: n = 21, items 1-7, 9, 10, 20-22, 24-26, 29, 31-33, 36, 39). The participants were requested to
talk the rater through the procedure, saying what they did or would do in a real OR setting on every consecutive step, thereby potentially allowing the raters to score all items, including those that could not be completed on the simulator (see Figure 1 for trainees at work).

All 3 faculty raters were practicing pediatric neurosurgeons, each of whom had extensive experience with neuroendoscopic surgery and were familiar with the steps involved in an ETV procedure. For each trainee, 2 of the 3 faculty raters graded each participant’s ETV performance live during the workshop. They received oral and written instructions on how to use the NEVAT procedural checklist and the GRS. Overall, they were encouraged to score the checklist as the participant was performing the procedure and score the GRS directly afterward. The final score was expressed as a percentage of items completed correctly on the checklist (total points/scored items) and as the mean score for the GRS (average of scores across all GRS items).

**Statistical Analysis**

Differences in scores between novices, experienced trainees, and experts were analyzed by Jonckheere’s trend test. If the outcome of Jonckheere’s test was
Validity evidence for NEVAT

significant, post hoc testing was performed with pairwise Mann-Whitney U tests with Dunn-Bonferroni correction. Internal consistency of the checklist and GRS were evaluated using Cronbach’s α with the 95% confidence interval (CI) plus the mean interitem correlation. The intraclass correlation coefficient (ICC) plus the 95% CI was reported as a measure of interrater reliability, based on data from the 2 raters who scored performances live during the workshop.

All data were analyzed using SPSS Statistics for Windows, Version 22.0 (IBM Corp, Armonk, New York) and R for Windows, Version 3.1.2 (R Foundation, Vienna, Austria) with statistical package “psychometric.” We considered P values < .05 as statistically significant. Missing data were taken care of by having calculations based on the mean total scores of the checklist and GRS. Descriptive demographic characteristics were reported as medians and interquartile ranges or frequency counts and percentages.

Results

Demographic Data

All faculty were neurosurgical consultants for whom ETV was a routine part of their clinical and teaching responsibilities (Table 1). For details on trainees, see Table 2. A total of 67 evaluations were completed (34 participants rated by 2 raters) of 13 faculty, 11 experienced trainees, and 10 novices. One missing GRS form (ie, all items of 1 GRS) belonged to a novice.

Internal Structure

High interrater reliability was observed between the 2 live raters, with the ICC for the checklist scores being 0.82 (95% CI: 0.63-0.91) and the ICC for the GRS

<table>
<thead>
<tr>
<th>Table 1. Demographics of experts (N = 13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N (% or median (IQR))</td>
</tr>
<tr>
<td><strong>Years as staff surgeon</strong></td>
</tr>
<tr>
<td><strong>ETV procedures performed</strong></td>
</tr>
<tr>
<td><strong>Prior exposure to simulation based training</strong></td>
</tr>
</tbody>
</table>

Abbreviation: IQR, interquartile range; ETV, endoscopic third ventriculostomy
being 0.94 (95% CI: 0.89-0.97). A moderate level of internal consistency was observed for the checklist, with Cronbach’s α = 0.74 (95% CI: 0.60-0.85) and mean interitem correlation = 0.066. The GRS had excellent internal consistency, with Cronbach’s α = 0.97 (95% CI: 0.95-0.98), with mean interitem correlation of 0.78.

### Relationships With Other Variables

The Jonckheere’s test revealed a significant trend in the data (J = 256, z = 2.04, P = .04, r = 0.35), with improving checklist scores observed as participant’s level of expertise increased; the trend did not remain when the corrected checklist was used with nonscorable items removed (J = 236, z = 1.41, P = .16, r = 0.24). Pairwise comparisons with adjusted P values showed no significant differences. Jonckheere’s test revealed a significant trend in the data, with the median GRS score improving with increased level of expertise (J = 291, z = 3.15, P = .002, r = 0.54) (Table 3). Pairwise comparisons with adjusted P values showed significant differences between novices and experts (P = .01). There was also a trend toward significance between experienced trainees and experts (P = .05); however, there was no significant difference between novices and experienced trainees (P = 1.0).
Validity evidence for NEVAT

Discussion

Simulation and Competency-Based Training
With current resident work-hour restrictions, mounting concerns regarding patient safety, and the shift toward competency-based methods of surgical training, the role of simulation-based training and assessment in postgraduate neurosurgical education is growing. Simulation provides a low-risk environment where trainees can practice and their performance can be assessed before they begin performing procedures on real patients. Although simulation has recently been applied in various aspects of neurosurgical training (eg, cerebral aneurysm clipping, intracranial tumor resection, pedicle screw placement, cervical laminoforaminotomy, and spinal durotomy), applications to neuroendoscopy may be particularly beneficial for improving patient safety. In the literature, several authors report a steep learning curve for ETV, with most errors occurring in the early stages of a surgeon’s experience with the procedure. However, simulation-based training has been noted to reduce surgical errors for other procedures and may demonstrate similar results when applied to ETV. These characteristics make simulation-based neuroendoscopic training an important target for simulation-based training assessment within the postgraduate neurosurgical curriculum.

To realize the potential of simulation and support the paradigm shift toward competency-based neurosurgical education, there is urgent need for standardized, procedure-specific methods for assessing surgical competence. The NEVAT was developed specifically for this purpose. In accordance with existing expert-based methods of assessment, the NEVAT relies on checklists and global ratings to evaluate a trainee’s technical skills. In addition to the content validity evidence previously generated by our research group on this tool, the current

Table 3. Median scores on checklists and GRS per group

<table>
<thead>
<tr>
<th>Group</th>
<th>Median checklist (IQR)</th>
<th>Median GRS (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novices (n = 10)</td>
<td>0.69 (0.58 – 0.86)</td>
<td>3.1 (2.5 – 3.8)</td>
</tr>
<tr>
<td>Experienced trainees (n = 11)</td>
<td>0.85 (0.63 – 0.89)</td>
<td>3.7 (2.2 – 4.3)</td>
</tr>
<tr>
<td>Experts (n = 13)</td>
<td>0.85 (0.81 – 0.91)</td>
<td>4.6 (4.4 – 4.9)</td>
</tr>
</tbody>
</table>

Abbreviations: GRS, global rating scale; IQR, interquartile-range
study generates evidence regarding the NEVAT’s relationships with other variables (ie, association with level of surgical expertise, in the classic validity framework known as construct validity) and internal structure (ie, interrater reliability and internal consistency).

Relationships With Other Variables
The validity evidence of the NEVAT with respect to its ability to distinguish between individuals at different levels of surgical expertise is demonstrated by the differences in scores observed among novices, experienced trainees, and expert surgeons performing ETV on a physical simulator. Even with our relatively small sample size, the GRS of the NEVAT was able to discriminate between novices and experts, and a trend toward significant differences between experienced trainees and experts was observed. There was also an overall significant trend in the median scores of the 3 groups that suggests that with increased level of training, GRS scores of onNEVAT improve, corroborating the relationship between NEVAT scores and surgical expertise. Furthermore, the experienced trainee cohort demonstrated the largest interquartile range, and their median score was closer to the novices than to the experts. As has been reported previously with other global rating instruments, these results suggest that the GRS of the NEVAT captures nuanced elements of surgical performance and as such may be particularly useful for assessment of more experienced trainees.

Conversely, on the procedure-specific checklist of the NEVAT, the median score of experienced trainees was closer to that of experts than that of novices. This finding is not surprising, as such checklists are designed to evaluate a trainee’s knowledge of the steps required to complete a procedure rather than nuances regarding surgical technique. It is possible that a plateau or ceiling effect among experienced trainees and experts was observed, as these individuals had likely already mastered the basic steps of ETV. Furthermore, the fact that the median checklist score was not close to perfect also highlights that, with experience, a surgeon may not always complete a procedure in a standard fashion, but in a more fluid manner that reflects variations based on their expert judgment of the case at hand. In turn, this supports the argument that procedural checklists are most useful for tracking learning among novices in the initial stages of training.
on a given procedure and is thus particularly useful for low-stakes formative assessment in which feedback is the goal.\textsuperscript{11} To facilitate this, the NEVAT checklist is divided into subsections, reflecting different parts of the ETV procedure. This allows educators to provide specific feedback and for trainees to focus on specific elements of the procedure in which they require further improvement (ie, deliberate practice).\textsuperscript{46}

**Internal Structure**

Our study also provides validity evidence regarding the internal structure of the NEVAT, given that the interrater reliability observed for both the checklist (0.82) and GRS (0.94) is comparable to published values for other checklist and GRS tools.\textsuperscript{45} In general, ICC values of ≥0.7 are considered acceptable, with values ≥0.9 desired for high-stakes assessment.\textsuperscript{47} As further validity evidence regarding internal structure, a high level of internal consistency of the GRS was demonstrated in our study. The observed Cronbach’s a value (0.97) and interitem correlation (0.78) exceeded published guidelines for minimum acceptable values, which typically range between 0.7 and 0.9 for Cronbach’s a and 0.15 to 0.5 for interitem correlations.\textsuperscript{47,48} However, as a values .0.9 may indicate that there is redundancy among items within the GRS, it is possible that some items of the GRS could be removed without losing much information from the assessment.\textsuperscript{47,49,50} As such, future work that tests shorter (and thus potentially more efficient) versions of the NEVAT GRS would be of value. Whereas Cronbach’s a of the checklist (0.74) is within acceptable values, the poor interitem correlation (0.066) suggests that this value may have been inflated by the large number of checklist items.\textsuperscript{48} One explanation for the poor interitem correlation observed in the checklist scores is that raters were given the option to mark an item with “not applicable.” In the synthetic simulator that is used for this study, it was not possible to perform certain aspects of the ETV procedure (eg, making an incision, placing the burr hole, closing the incision). To account for this, participants were asked to talk the raters through the procedure while they were performing an ETV. If a step of the procedure could not be physically performed, the participant would say what action normally would be taken. This
may have led to a situation in which some raters scored “not applicable,” whereas others scored the item based on a candidate’s description of the step in question. Further work to clarify the consistency of checklist ratings during operative cases (in which the full checklist can be assessed) may address this issue.

Limitations and Future Work

There are some limitations to this study. First, the small sample size decreased the power in our study to differentiate between performances of novice, experienced, and expert participants.

An additional limitation is the number of items in the checklists and GRS. Although reducing the number of items may improve usability and feasibility of these tools (thereby improving their uptake in both the clinical and simulated setting), such item reduction may come at a cost because the ability to discriminate between individuals at different levels of experience might decrease. Finally, as contemporary validity frameworks view validation as an argument supporting the use of an assessment instrument in a specific population for a specific purpose, based on the current data, the validity of the NEVAT cannot be assumed to extend beyond the simulated setting (ie, to assessment of performance in the OR) or to other neuroendoscopic procedures other than ETV. Future work to generate validity evidence related to response process (eg, think-aloud protocols of expert raters as they complete the assessment to determine what information they are using to generate their ratings) will also be useful to ensure that scores generated using the NEVAT accurately reflect surgical competence in ETV. The NEVAT could eventually be used as a standard evaluation method especially in the context of competency-based education. There is a need for ongoing assessment of trainees’ performance at various stages along the educational continuum. Based on the validity evidence generated in this study, the NEVAT can now be used for simulation-based assessment of ETV for both novice and experienced learners. These tools could also be used to assess the effectiveness of different simulation platforms or curricular designed for ETV training. As was previously proposed by Hooten et al, residents could also be required to demonstrate a certain level of competency (ie, achieve certain scores on the NEVAT) in a simulation-based setting before performing an ETV on a patient. Finally, the NEVAT also has the potential to be useful as an assessment tool for
postgraduate credentialing; however, stronger validity evidence (eg, related to response process and consequences) will need to be generated to justify its use in such high-stakes assessments.

**Conclusion**

We generated validity evidence for a series of assessment instruments designed to evaluate neuroendoscopic competence in the ETV procedure by reporting on the internal structure and its relationships with other variables during simulation-based training. Additional validation studies need to be conducted because the process of validation is not a one-time event. The development and evaluation of assessment tools such as the NEVAT can support the paradigm shift toward competency-based education by providing a method for standardized and transparent evaluation of surgical competence for neuroendoscopy training.

**Acknowledgments**

The authors thank all faculty and trainees of the neuroendoscopy workshop for their participation, and specifically Dr Spyros Sgouros and Dr Piero Spennato for their help with the live rating of the participants. They also thank Dr Santolo Cozzoline, Director of the Center of Biotechnologies of the Cardarelli Hospital, for logistic support of this study.
CHAPTER 5

References


Validity evidence for NEVAT


CHAPTER 5

Validity evidence for NEVAT