

Development of Hand Skills Testing Methodology for O&P Curriculum

Meghan Roach, MSOP, CPO, Jeremy Sherman, MSOP, CPO

ABSTRACT

Introduction: In orthotics and prosthetics education, there is a need for objective means of tracking and assessing hand skills development. These specialized skills are essential for clinicians when fabricating and modifying patients' devices but are not learned in the same manner as theoretical knowledge. This research is the foundational step in designing and implementing a Hand Skills Test (HST) methodology and rubric within the context of the Baylor College of Medicine's Orthotics and Prosthetics program. This study addresses the challenge of objectively evaluating students' hand skills acquisition and competency, which is a key component of clinical education.

Methods: Trials were administered at the onset and conclusion of participants' first year at the program. Five tasks common in the field were selected: metal bending, riveting, plastic flaring, foam skiving, and plastic polishing. Rubrics were designed to assess the task products. After reviewing the grading process, revisions were made to increase efficiency, testing content validity, and interrater reliability.

Results: Rubric revision increased grader percent agreement minimally. Percent agreements for all evaluations were in the upper 80s, which is within the acceptable range. Participants' HST minimum scores increased, and the variance of scores decreased scores between trials 1 and 2. The results follow expected trends, lending credibility to this assessment tool.

Conclusions: This research is the preliminary phase of the study, which will collect testing data for the Baylor College of Medicine's Orthotics and Prosthetics program cohorts through 2029. Future investigations will examine trends in hand skill development during participants' time at the program. The goal of this project is to create an assessment tool to help identify how these specialized skills are learned and refined.

Clinical Relevance: By better assessing and understanding the acquisition of hand skills, orthotics and prosthetics clinical education can be enhanced to best meet learners' needs. (*J Prosthet Orthot.* 2025;37:148–152)

KEY INDEXING TERMS: hand skills, curriculum development, O and P education

INTRODUCTION

In graduate level clinical education for orthotics and prosthetics (O&P), much of the training takes place in the classroom, studying anatomy, biomechanics, and the principals behind O&P devices. In addition to this theoretical knowledge, clinicians working in the field are expected to fabricate or modify these devices, necessitating the teaching of requisite hand skills. The principles and methods needed to perform these skills safely can be discussed in the classroom, but mastery is obtained by combining that understanding with observation and practice.¹ Assessing learners' hand skill

competency is difficult due to its subjective nature, yet necessary in education.

The Commission on Accreditation of Allied Health Education Programs (CAAHEP) lists 20 foundational content areas required for master's programs in O&P, one being these hands-on "technical skills." This is defined in part as "the study and supervised practice of the psychomotor skills necessary to ensure safe and appropriate use of tools and equipment to formulate and implement orthotic/prosthetic treatment plans."² At Baylor College of Medicine's Orthotic and Prosthetic program, hand skills are taught throughout two semesters of Technical and Safety Skills (TSS) classes.

There have been many studies assessing the best methods for teaching psychomotor skills in various health care fields.^{1,3} Understanding effective training methods for teaching these unique types of skills is imperative, but equally important is using standards to measure the acquisition and competency of these skills. Having a way to monitor O&P hand skill acquisition and determine competency means establishing a system to objectively evaluate and track learners' progress.

Other health care fields have also developed curriculum around the combined education of theoretical knowledge and psychomotor skills. Research in these other fields looks at factors that play a role in predicting the success of trainees and how to rate competency.^{4,5} The fields of dentistry, physical therapy, and

MEGHAN ROACH, MSOP, CPO; and JEREMY SHERMAN, MSOP, CPO, are affiliated with Baylor College of Medicine, Houston, Texas.

Disclosures: The authors declare no conflict of interest. Authors state no funding was received.

Copyright © 2024 American Academy of Orthotists and Prosthetists.

Correspondence to: Meghan Roach, MSOP, CPO, Baylor College of Medicine, 1 Baylor Plaza, MS 115, Houston, TX 77030; email: meroach@gmail.com

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (<https://journals.lww.com/jpojournal/pages/default.aspx>).

surgery have developed simulations, virtual reality assessments, and task-specific tools to objectively rate trainees' hand skill learning.⁶⁻⁹ Our goal with this research is to establish a similar assessment tool for O&P education.

METHODS

STUDY PARTICIPANTS

This study was approved by the Institutional Review Board at Baylor College of Medicine.

Participants in this study are students at BCMOP, in the graduating class of 2024. Inclusion criteria were as follows:

- Current BCMOP students.
- Enrolled in either TSS I or II.
- A minimum of 1 month of classroom instruction in hand skills fundamentals for safety.
- No previous technical experience is required for program enrollment, although 150 hours of clinical shadowing are suggested for program applicants.

Students could decline from being included in the study, although they were still required to complete the testing as a part of TSS curriculum. Additionally, per laboratory safety rules, students were advised that they may opt out of any task if they do not feel comfortable using required machinery.

TSS curriculum introduces, educates, and evaluates technical skills necessary for completion of core O&P competencies. Demographic data about students' prior O&P experience have not yet been collected; however, there are plans to do so in the future to help analyze baseline performance variation.

TRIAL 1

The BCMOP curriculum is made up of 12 months in the classroom, followed by 18 months of clinical residency. Testing for this study takes place twice during those first 12 months. Trial 1 occurs in the 2nd month, and trial 2 occurs in the 12th month. Each trial consists of the O&P Hand Skills Test (HST), which was formalized for the purpose of this research and administered as part of BCMOP's TSS I and TSS II classes.

Prior to trial 1, students have been introduced to HST tasks, with an emphasis on safe completion of fundamentals, but have not had time to refine skills through practice or self-study. Trial 1 establishes baseline scores before prolonged exposure to hand skills during TSS. Trial 2 happens after the year of onsite education, training, and repetition of hand skills. During this year, students regularly utilize hand skills while completing the projects within their core courses. Students receive feedback on the craftsmanship and functionality of their projects from BCMOP faculty and patient models.

TESTS

The HST is a formalized assessment derived from previous TSS evaluation tools. Tasks included in the HST are meant to capture and objectively assess the execution of tasks frequently used by O&P clinicians. Many hand skills fit this criterion; however, those

included in this study had to be easily repeatable and utilize dependable equipment. The initial tasks selected were metal bending, riveting, plastic forming, foam skiving, and plastic polishing.

Documents were created for each task listing and illustrating materials/equipment, safety considerations, steps, and ideals for the finished product (see Document, Supplemental Digital Content 1, <http://links.lww.com/JPO/A144>). Documents are emailed to participants 1 week in advance of testing, reviewed just prior to the trial, and are provided at each station during the HST. Participants are asked to read the instructions and reference documents as needed; however, familiarization with the content is up to individuals.

HST takes place in the O&P laboratory at Baylor College of Medicine, with stations set up for each task. Stations are equipped with all necessary tools, equipment, documents, and an example of the finished product. HST trials start with gathering participants for instructions. Task documents and explanation of station rotation and timekeeping are read aloud (see Document, Supplemental Digital Content 2, <http://links.lww.com/JPO/A145>). Boxes are issued to each person, containing necessary task materials, a pen, and a timecard (see Document, Supplemental Digital Content 3, <http://links.lww.com/JPO/A146>). Boxes are labeled with a letter, and proctors record participants' names with the corresponding letters.

A proctor with a stopwatch is stationed in the area. Eight minutes are allotted for completion of each task. If a participant finishes early, they self-report their time on the timesheet. If they do not complete a task in the given time, "incomplete" is marked on the timesheet. Once the HST is complete, task products are labeled with the box letter, grouped by task, and are kept in the locked laboratory office until grading takes place. Timesheet data are deidentified and entered into a database by the researchers.

Three existing validated tests are also conducted along with the HST in each trial:

- The Purdue Pegboard Test (assessing manual dexterity)
- Grip strength testing using a dynamometer
- The Revised Purdue Spatial Visualization Test (PSVT-R)

These tests were chosen based on their perceived relevance to the chosen O&P hand skills, as well as their inclusion in testing for similar technical professions.¹⁰⁻¹² Testing is done in accordance with manufacturer or designer written standards. At this time, data have been collected with plans for correlation analysis in the future.

GRADING

Separate rubrics were created for each of the five tasks to assess the products' craftsmanship and accuracy to instructions. Rubrics analyzed between 8 and 11 characteristic items. Each task product was assigned an item score by three different graders, and those were tallied to give an overall score from each grader. Graders were certified clinicians and BCMOP faculty who were educated on rubric use prior to grading. They had paper scorecards that were collected by the researchers, and the data were entered into the

spreadsheet. After initial grading of trial 1, graders were asked for their input on the process.

RUBRIC REVISION

When assessing technical skills, Aggarwal highlights the importance of not only reliability and objectivity, but ease of performing the evaluation.¹³ Graham builds on that to say that a rubric's efficiency can impact the consistency of the grader.¹⁴ The intended scope of future research using the HST and grading rubrics is 5 years, with trials eventually taking place three times per year. For these reasons, we sought the graders' feedback to improve the process. Ease of the process is critical for continued participation of graders. According to their input, time involvement was the biggest concern. Although time spent grading varied among the graders, it was not uncommon to exceed 30 minutes to grade only one of the total five tasks. There was also feedback that the paper scorecard system was not straightforward since some items required nuanced grading on a scale of quality, and the graders preferred a binary 0–1 grading metric for each task for ease and speed.

SYSTEMATIC REVIEW

Other factors that were examined when editing the rubric were interrater reliability and internal consistency. Interrater reliability was examined by calculating grader percent agreement (PA). Individual rubric items with especially low agreement were reviewed by an expert panel. Interitem correlation was used to determine if multiple items on the rubric were related to each other (for instance, if one item was checked yes, then the second item was also always checked yes, meaning that they did not grade distinct measures of quality). If two items had high correlation, they were combined, thus reducing redundancy. Cronbach's alpha also examines internal consistency and determines how well items measure task characteristics. If internal consistency could be improved by removing an item, it was reviewed and possibly deleted based on expert opinions. Figure 1 diagrams this process.

Revisions to the rubric were done with four primary goals: reduce grading time, increase grader clarity, increase interrater reliability, and improve internal consistency. Based on the expert opinion reviews, some items were deleted, combined, and reworded. Items were converted to a dichotomous yes/no format, and the rubric was changed to an online format. Addition-

ally, the number of graders was increased to 4 to reduce interrater agreement by chance. The revised rubric is available to view in the additional materials accompanying this paper (Document, Supplemental Digital Content 4, <http://links.lww.com/JPO/A147>).

TRIAL 2 (CHANGES TO TESTING)

In addition to rubric revisions, changes were made to the testing process to improve efficiency for participants and researchers. Modifications were made using trial 1 insights, while being sensitive to the integrity of the HST. The metal bending task was eliminated for several reasons:

- The skill is not taught until semester 2, so there is no introduction prior to trial 1.
- Only one curriculum project features this technique, so there is less opportunity to practice and improve.
- Metal bending is a complex skill, and while it is covered by CAAHEP standards, in a clinical setting, it is usually executed by experienced technicians (not clinicians).
- Eliminating metal bending focuses the HST on more critical hand skills for clinicians and cuts down on overall grading time.

Other changes to the HST included increasing station time from 8 minutes to 10 (which meant overall testing time per student remained 40 minutes). Also, machinery is available to have two participants working simultaneously at each station, cutting down on testing time.

RESULTS

Interrater reliability was examined from trial 1 grading with the initial rubric, and those data were used to systematically edit the grading process to produce a revised rubric. The same measures were used to compare grading of trial 1 and trial 2 with the revised rubric. When independently assigning a score for an item, ideally all graders would be in complete agreement; however, this is improbable. Agreement was examined by calculating PA between graders for each item of each task on the rubrics, as well as overall task PA, and an overall HST Trial PA. Literature suggests that PA between 75% and 90% is acceptable, with the fewer rating levels in a rubric, the higher the acceptable PA should be.^{14,15} With our initial rubric being mostly dichotomous, and the revised rubric being entirely dichotomous, we wanted our PA to be as high as possible.

Trial 1 overall HST PA with the initial rubric was 86.1%. Although this is well between the 75% and 90% range, we chose to rework the rubric for the reasons described in the methods section, with one objective being to increase PA. After revising the rubric and regrading trial 1, the overall HST PA went up slightly to 88.4%. Trial 2 was only graded with the new rubric, and the overall HST PA was 87.2% (see Table 1).

Standard deviation was calculated per grader, per task, to examine the variance of scores within a trial. Between trials, standard deviation decreased 75% of the time, indicating less variance in graders' overall task scores for trial 2 (see Table 2). When looking at the lowest (minimum) overall scores assigned by

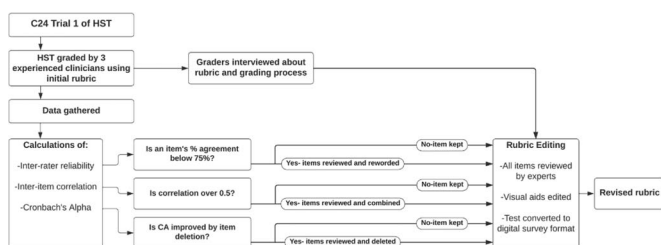


Figure 1. The systematic process for improving the initial rubric, using grader feedback and calculation of interrater reliability and internal consistency, with the intention of creating a more efficient and reliable evaluation process.

Table 1. Graders’ percent agreement for trial 1 increased slightly after the rubric was reworked.

	Trial 1 Initial Rubric	Trial 1 Revised Rubric	Trial 2 Revised Rubric
Riveting	89.9%	92.2%	92.0%
Flaring	87.1%	91.8%	91.1%
Skiving	85.8%	87.7%	83.8%
Polishing	81.5%	83.7%	81.9%
Overall HST	86.1%	88.9%	87.2%

Percent agreements for all three grading attempts are within 2.8% of each other, indicating minimal impact of the rubric revisions, bolstering the means of our assessment.

graders per task, 81% of the time the lowest assigned score for trial 2 was higher than the lowest score assigned for trial 1.

Individual participants’ scores between trials 1 and 2 can be compared using the revised rubrics scoring. This is something that future researchers will be looking at in depth once more data have been collected. For now, we made simple calculations to help us understand the differences between the rubrics. For the overall HST, 65% of participant’s task scores improved between trial 1 and trial 2. Between the trials, class averages changed as follows: riveting decreased 2.09%, flaring had a 0.76% increase, skiving 20.21% increase, and polishing 27.92% increase (Figure 2). This translates to an overall HST improvement of 11.70% for the class of 2024.

DISCUSSION

The HST is only beneficial as a means of objective skill acquisition assessment if grading is consistently repeatable between trials and graders. In reworking the rubric, one of the objectives was to increase the PA to make it more reliable. The difference between the initial and revised rubric increased 2.8% for trial 1. Although this is a slight improvement, both PAs plus the trial 2 grading are all well within the acceptable window, indicating an appropriate degree of interrater reliability. PA similarity indicates that the revisions did not significantly alter the assessment criteria, reinforcing the consistency of our evaluation approach.

Table 2. The four graders’ standard deviations difference between trial 1 and trial 2 scoring are shown per task.

	Flaring	Polishing	Skiving	Riveting
G1	−0.89	−0.24	−0.24	−0.74
G2	−0.60	0.19	0.24	−0.38
G3	−0.36	−0.12	0.08	0.03
G4	−0.51	−0.05	−0.45	−0.69
Avg	−0.59	−0.06	−0.09	−0.44

As expected, average standard deviation decreased for all four tasks, indicating less score variation after participants had a year of on-site instruction and practice versus the baseline scores.

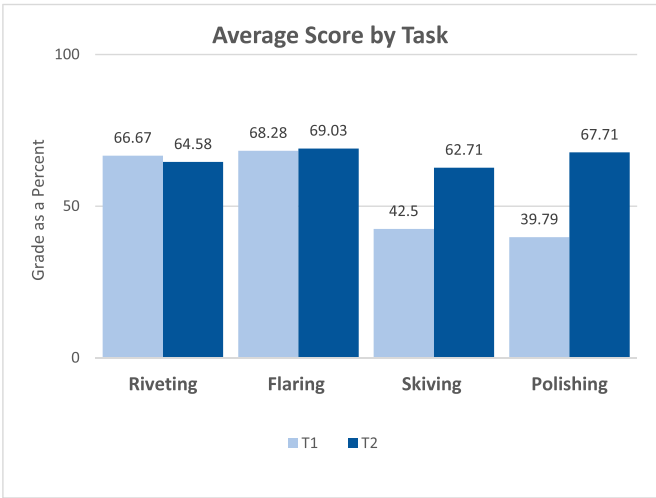


Figure 2. The class of 2024 participants’ HST average grades were compared between entering the program and after practicing skills during 1 year of onsite training. Scores increased greatly for the polishing and skiving tasks and changed minimally for the flaring and riveting tasks.

One possible explanation for PA improvement not being greater is that the initial rubric had a total of 45 items from five tasks, whereas the revised rubric had only 30 items from four tasks. This change was to make grading more efficient; however, reducing items by one third creates fewer opportunities for interrater agreement. Additionally, the initial rubric was graded with three graders, and all subsequent grading took place with 4. The likelihood of agreement further decreases with more people assessing fewer items, perhaps explaining the minimal increase in the PAs. Nonetheless, our findings support the credibility of our rubrics and their applicability in the educational context.

The decrease in graders’ standard deviation and increase in minimum scores between trials are anticipated trends for students learning a new skill, which adds confidence in the accuracy of our assessment tool. It is likely that participants’ scores would have high variation when testing baseline because of their diverse hand skills background entering into the program. The variance should narrow once skills are practiced and honed throughout the year. Accordingly, minimum scores should improve along with overall average scores. The initial testing and grading follow expected trends, which lends credibility to the HST and rubric.

Common sense suggests individual participants’ scores should improve between trials because of increased exposure and practice. This assumption is backed up with nearly a 12% overall improvement in scores after 12 months in TSS classes. This is skewed, however, with the majority of higher scores coming from the polishing and skiving tasks. One possible explanation is that participants spend the most time honing these skills over various projects, so greater improvement was shown. Having the data from only one cohort, it is premature to say if this is a trend, and thorough analysis of this will be left to future research.

LIMITATIONS

The aspects of assigning scores to hand skills samples are inherently subjective, with the risk that scoring reflects graders’

preference instead of participant performance.¹⁴ Creating rubrics, educating graders, and systematically making rubric revisions maximize grader's objectivity.¹⁵ Our high PA scores reflect this. Percent agreement has long been used as a means of assessing interrater reliability. When contrasting means of calculating interrater reliability, McHugh concedes that while easy to calculate and interpret, this measurement tool may be oversimplified because it does not account for graders assigning same scores by chance.¹⁶ The likelihood of agreement by chance is increased because of our dichotomous revised rubric. This grading scale was employed for ease and efficiency of grading; however, it may "restrict evaluators' choices to the detriment of accuracy."¹⁴ Cohen's kappa and other kappa statistics have been developed to factor in that uncertainty and are a valued tool for evaluating interrater reliability.^{14,16} These statistics were not appropriate for assessing data from our rubrics because of the binary structure, nonmutually exclusive ordinal nature, and our use of three or more raters.

CONCLUSIONS AND FUTURE DIRECTIONS

This research is the first step in a project to study and track hand skill development of students at BCMOP throughout their time in the program. The goal of this phase was to develop a reliable and repeatable assessment tool. With only the first year of testing complete, the value of the instrument is yet to be determined; however, we argue that our revision process for tests and rubrics by an expert panel using grader input increases our content validity and interrater reliability. We believe that this preliminary use of our assessment tool measures what it was designed to measure.

With the research continuing through 2029, future research will explore 5 years of HST data for trends over time and within each class. HST scores will also be compared with the data collected from the existing validated tests, which are administered concurrently, as well as a demographic survey to explain initial skill disparities. They will be able to study data for correlation among participants' background, experience, manual dexterity, spatial awareness, grip strength, and performance on the HST. The overarching aim of our research is to understand how these specialized skills are learned over time, in order to better inform the education of O&P hand skills.

REFERENCES

- Nicholls D, Sweet L, Muller A, Hyett J. Teaching psychomotor skills in the twenty-first century: revisiting and reviewing instructional approaches through the lens of contemporary literature. *Med Teach* 2016;38(10):1056–1063. doi:10.3109/0142159X.2016.1150984.
- Standards and Guidelines for the Accreditation of Educational Programs in Orthotics and Prosthetics Essentials/Standards Initially*
- Causby RS, Reed L, McDonnell MN, Hillier SL. Teaching of Manual Clinical Skills in Podiatric Medicine Theory and Recommendations. *J Am Podiatr Med Assoc* 2018 Mar;102(2):158–167. doi:10.7547/15-223. PMID: 29634305.
- Gallagher AG, Leonard G, Traynor OJ. Role and feasibility of psychomotor and dexterity testing in selection for surgical training. *ANZ J Surg* 2009;79:108–113. doi:10.1111/j.1445-2197.2008.04824.x.
- Psychomotor ability testing and human reliability analysis (HRA) in surgical practice. *Minim Invasive Ther Allied Technol* 2001;10(3):181–195. doi:10.1080/136457001753192312.
- Beard JD, Jolly BC, Newble DI, et al. Assessing the technical skills of surgical trainees. *Br J Surg* 2005;92(6):778–782. doi:10.1002/bjs.4951.
- Schreiber J, Gagnon K, Kendall E, et al. Development of a grading rubric to assess learning in pediatric physical therapy education. *Pediatr Phys Ther* 2020;32(1):70–79. doi:10.1097/PEP.0000000000000667.
- Segura C, Halabi D, Navarro N. Design and validation of a basic dental psychomotor skills test for novice dental students. *J Dent Educ* 2018;82(10):1098–1104. doi:10.21815/jde.018.111.
- Lugassy D, Levanon Y, Shpack N, et al. An interventional study for improving the manual dexterity of dentistry students. *PLoS One* 2019;14(2):e0211639. doi:10.1371/journal.pone.0211639.
- Lugassy D, Levanon Y, Pilo R, et al. Predicting the clinical performance of dental students with a manual dexterity test. *PLoS One* 2018;13(3):e0193980. doi:10.1371/journal.pone.0193980.
- Constansia RDN, Hentzen JEKR, Buis CI, et al. Is surgical subspecialization associated with hand grip strength and manual dexterity? A cross-sectional study. *Ann Med Surg (Lond)* 2021;73:103159. doi:10.1016/j.amsu.2021.103159.
- Maeda Y, Yoon SY, Imbrie PK, Kim-Kang G. *Psychometric Properties of the Revised PSVT: R for Measuring First Year Engineering Students' Spatial Ability Psychometric Properties of the Revised PSVT:R for Measuring First Year Engineering Students' Spatial Ability**; 2013. <https://www.researchgate.net/publication/261760571>
- Aggarwal R, Grantcharov TP, Darzi A. Framework for systematic training and assessment of technical skills. *J Am Coll Surg* 2007;204(4):697–705. doi:10.1016/j.jamcollsurg.2007.01.016.
- Graham M, Milanowski A, Miller J, Henderson A. *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings*; 2012.
- Stemler SE. A comparison of consensus, consistency, and measurement a comparison of consensus, consistency, and measurement approaches to estimating interrater reliability approaches to estimating interrater reliability. *Pract Assess Res Eval* 2004;9:4. doi:10.7275/96jp-xz07.
- McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;22:276–282.

Adopted in 1993; 2001. <https://ncope.org/wp-content/uploads/2022/01/Residency-Standards-Dec-2021-Revision.pdf>