Development and Validation of Tools for Evaluation of Orthosis Fabrication

Andonia Stefanovich, Camille Williams, Pat McKee, Eric Hagemann, Heather Carnahan

This study is the first phase of research aimed at developing new educational approaches to enhance occupational therapy students' orthosis fabrication skills. Before the effectiveness of training can be determined, a method for evaluating performance must be established. Using the Delphi method, we developed a global rating scale and checklist for evaluating technical competence when fabricating metacarpophalangeal (MCP) joint-stabilizing orthoses. To determine the reliability and validity of these tools, three evaluators used them to assess and score orthotic fabrication performance by experienced and student occupational therapists. The results suggest that these measurement tools are valid and reliable indicators of the technical skills involved in fabricating an MCP joint-stabilizing orthosis. Future studies should focus on building on these tools to evaluate communication skills, technical skills for making other types of orthoses, and effectiveness of training programs.


Upper-limb orthoses (splints) made from low-temperature thermoplastics are commonly fabricated by occupational therapists; as such, orthotic education is a standard part of the curriculum in most occupational therapy programs (Clark, 2002). Occupational therapy students are taught this skill through the traditional teaching methods of didactic lecture-based learning, readings, observation, and hands-on practice with fellow students who often have limbs without any pathological deformities. This situation is problematic because occupational therapy students are not given the opportunity to learn and practice the skills required to construct orthoses for people with pathological deformities and because no standard, valid, and reliable measurement tool is available to evaluate students’ technical competence at creating such orthoses. As a result, the performance of new occupational therapists may be subject to a learning curve; the consequences of this learning curve for clients are unclear.

One solution to moving students further along on the learning curve before they deal with real clients is simulation. Lammers (2007, p. 505) defined simulation as “the artificial representation of a situation, environment, or event that provides an experience for the purposes of learning, evaluation, or research.” Simulation training allows students to practice and operationalize new knowledge and transform concepts into practical skills (Lammers, 2007). Simulation training has been used successfully for decades to train professionals in various fields, including dentistry, nursing, aviation, and medicine, and it is becoming an educational standard for teaching technical skills to trainees in these fields. Students learn and practice technical skills on models and simulators to better prepare to perform such skills in the clinical environment. Binstadt et al. (2007) found that several studies have demonstrated improvements in the performance of
The purpose of this study was to develop and validate measurement tools to evaluate the technical skills involved in the creation of an MCP joint–stabilizing orthosis. We chose this orthosis because it is less complicated to cut and mold than other commonly encountered orthoses and is also suitable for fabrication on a simulated hand. A Delphi survey with experts in orthotic fabrication was conducted to develop the measurement tools. The Delphi method is a systematic method for gathering and organizing information and opinions from a panel of experts on a complex issue or problem without having the experts meet physically (Vázquez-Ramos, Leahy, & Hernandez, 2007). The Delphi methodology was developed in the 1950s for use in the military but has since been used in many fields, including health professions education (de Villiers, de Villiers, & Kent, 2005; Palarca, Johnson, Mangelsdorff, & Finstuen, 2008). The technique has evolved over time, producing several variations (Crisp, Pelletier, Duffield, Adams, & Nagy, 1997); however, the goal and general process have remained the same: The investigators identify a question or issue and then generate an initial list of items to address the question or issue that is submitted electronically and iteratively to the members of the Delphi panel, who provide their expert opinions. Opinions are then analyzed and refined iteratively through feedback, building consensus by enhancing the individual opinions of experts (de Villiers et al., 2005; Graham, Regehr, & Wright, 2003). The process has the advantage of allowing professionals to participate in an otherwise intensive process by reducing the time and resources required to meet physically. Panelists remain anonymous so that the group process is not unduly influenced by the reputation or opinion of any one panelist (Graham et al., 2003).

Subsequent to development of a GRS and checklist, we conducted validation tests to determine whether the tools were suitable for evaluating occupational therapy trainees learning to fabricate orthoses. We hypothesized that participants with more experience (practicing occupational therapists) would obtain higher scores than novices (occupational therapy students).

Development of the Measurement Tools

Participants in the item generation phase were required to have at least 1 yr of experience as an occupational or hand therapist but not necessarily experience fabricating MCP joint–stabilizing orthoses. Two therapists were recruited from the University of Toronto community; each therapist had >20 yr of experience in hand therapy. The therapists independently generated items for the GRS and checklist using a photo of an MCP joint–stabilizing
orthosis, a sample surgical GRS and checklist as a guide and formatting reference, and their clinical knowledge and expertise. The checklists and GRSs developed by the therapists were then combined to create the initial checklist and GRS items.

For the item evaluation phase of the Delphi process, we prepared a questionnaire that would allow the Delphi panelists to rate the importance of each item on the initial list as an indicator of the technical performance of people creating an MCP joint–stabilizing orthosis. Research has demonstrated that the range for an optimal number of alternatives on a rating scale is between four and seven (Lozano, Garcia-Cueto, & Muniz, 2008), so our questionnaire used 7-point Likert scales (1 = not at all important, 7 = very important) for rating the importance of each item. Inclusion criteria for the Delphi panelists in this phase were similar to those for the item generation phase. Three new therapists from the Toronto Hand Interest Group agreed to participate; these participants had 3–19 yr experience in hand therapy. Each panelist was sent an e-mail with the initial list of items and the questionnaire for rating the importance of the items. A positive consensus was defined as at least two of the three panelists agreeing on an item by assigning it a score of ≥5 on the 7-point Likert scale. Consensus was reached for both measurement tools after this round of the Delphi process, so no further rounds were necessary.

Final consensus yielded a 7-item process-based GRS and a 15-item, task-specific, process- and product-focused checklist. The items on the GRS were respect for tissue, handling of scissors, handling of thermoplastic during heating and molding, flow of orthotic fabrication, knowledge of orthotic fabrication, overall performance, and quality of final product. To enhance ease of use, each item was scored on a 5-point Likert scale with explicit descriptive anchors at 1, 3, and 5, where 1 = lowest level of performance and 5 = ideal performance. For example, the anchors for the handling-of-scissors item were “repeatedly makes tentative or awkward cuts with scissors resulting in very rough edges,” “competent use of scissors but occasionally appeared stiff or awkward or created some rough cuts,” and “fluid movement with scissors resulting in smooth edges.” The total score for the GRS is the sum of the scores for the 7 items for a maximum total score of 35 points.

Two of the 5 process-focused items on the checklist were “appropriate choice of thermoplastic” and “checked temperature of heated thermoplastic before application to mold.” Two of the 10 product-focused items were “no restriction of the wrist” and “no redness of the skin indicating pressure points.” On the 15-item checklist, each item was awarded 1 point if it was completed correctly and no points if it was incomplete or incorrect, for a maximum total score of 15 points. (We encourage interested readers to contact Heather Carnahan to obtain complete copies of the final assessment tools.)

Method

Participants

Fifteen participants (6 experienced occupational therapists and 9 novices) were recruited to take part in the validation phase of this study. Experienced occupational therapists were defined as those having at least 1 yr work experience in hand therapy or more than 5 yr experience working in an area in which orthotic fabrication was a requirement. Although experience in fabricating orthoses was required, participants did not have specific experience creating an MCP joint–stabilizing orthosis. Therapists’ work experience ranged from 3 yr to 20 yr.

Novices were defined as occupational therapy students who had no previous experience in orthosis construction, including but not limited to fieldwork placement experience, volunteering, and previous job experiences. Novices were recruited from the 1st-year occupational therapy class of the University of Toronto. The institutional Office of Research Ethics approved the tool validation protocol, and all participants provided voluntary informed consent before participating, in accordance with the guidelines set out by the 1964 Declaration of Helsinki (World Medical Organization, 1996) and the institutional research ethics board.

Procedure

All data collection took place in the orthotics laboratory of the institutional Rehabilitation Sciences building. We adapted the fabrication process for the MCP joint–stabilizing orthosis so that the strapping component was excluded to reduce the time commitment involved and to simplify the design such that participants at all levels of training could be included. Participants viewed a previously created training video demonstrating the fabrication of an MCP joint–stabilizing orthosis, which is used as part of orthosis fabrication education in the institution’s occupational therapy curriculum. Participants were videotaped while creating the orthosis on a live nonpathological human hand using 2.4-mm-thick Aquaplast® Thermoplastic (Sammons Preston Canada Inc., Mississauga, ON). Videotaped performances captured only the participants’ hands and arms, not the head or face, to ensure that participants were not identifiable.

Three raters, blinded to the participants’ level of experience, watched the videotapes, examined the final orthoses
(Figure 1), and used the GRS and checklist to assess the participants’ technical skills with respect to the fabrication process and the final product. The raters had different levels of experience with hand therapy and orthosis fabrication. The rater group consisted of (1) a certified hand therapist with 23 yr experience in hand therapy and who was involved in continuing education, (2) a licensed occupational therapist, and (3) an informed nonclinician (graduate student in rehabilitation science) familiar with the task of creating orthoses through research in the area of education and clinical skills. All raters were instructed to review the measurement tools and view the training video.

**Data Analysis**

*Interrater reliability* is an important psychometric property that refers to the consistency of measurements between independent raters (Zeller, 1990). It ensures that different raters can use the tool to evaluate learners and obtain consistent or similar results. We examined interrater reliability for each tool by comparing scores from each rater for the same performances using single-measures intraclass correlation coefficients (ICCs) with a 95% confidence interval (CI). The two-way random effects model with both absolute agreement and consistency methods was used. The absolute agreement method determines whether the raters assigned similar scores (absolute values) for similar performances, and the consistency method determines whether the raters’ scores followed similar trends for the performances even if the absolute scores were not the same. For the type of assessment for which these tools will be used (formative or summative classroom-type assessment), acceptable reliability has been suggested to be ≥.70 (Downing, 2004).

*Construct validity* is the tool’s ability to measure the construct that is intended—in this case, skill in creating an MCP joint–stabilizing orthosis—and can be inferred if the assessment is able to discriminate between distinct levels of skill or expertise (Thorndike, 1990). We assessed the construct validity of the GRS and checklist by comparing the scores from each rater for novice (n = 9) and experienced (n = 6) therapists. Comparisons were made using a mixed-model analysis of variance with rater (hand therapist, licensed occupational therapist, informed nonclinician) as a within-subject variable and group (novices, experienced therapists) as a between-subjects variable. Post hoc pairwise comparisons between raters were analyzed using Bonferroni adjustments as calculated by SPSS Version 17.0 (SPSS Inc., Chicago). Briefly, the adjustment controls the familywise error by correcting the level of significance to α divided by the number of pairwise comparisons (.05/3 = .016 for this study). Statistical results were considered significant at p < .05. To help determine the importance of group main effects, Pearson’s correlation coefficient r effect sizes were calculated using the formula

\[ r = \sqrt{\frac{F(1, df_R)}{F(1, df_R) + df_R}} \]

where \(F(1, df_R)\) is the F ratio of the effect and \(df_R\) is the residual or error degrees of freedom. Effect sizes were

---

**Figure 1.** Pictures of metacarpophalangeal joint–stabilizing orthoses created by (A) an experienced occupational therapist and (B) a novice occupational therapy student.
considered small, medium, and large at .10, .30, and .50, respectively (Cohen, 1988, 1992, as cited in Field, 2009, p. 57). Analyses were done using SPSS Version 17.0.

**Results**

We conducted a reliability analysis using scores for experienced therapists and novices from three different raters. As seen in Table 1, the ICC values indicated that there was moderate agreement and consistency between raters for each measurement tool. Overlap between CIs also suggests that no significant differences were found in agreement or consistency between the measurement tools. Interestingly, ICC values for both consistency and agreement increased when all scores (global rating and checklist scores) were used to compute the ICC. Because these ICC values were very close to the threshold for acceptable correlation (moderate) and the lower bound of the CIs fell below the level of acceptable correlation, we did not average scores across raters when evaluating construct validity.

Analysis of the GRS scores indicated no main effect of rater, \( F(2, 26) = 0.83, p = .45 \), but a main effect of group, \( F(1, 13) = 16.13, p = .001, r = .74 \) (see Figure 2). On average, the experienced therapists were rated significantly higher (mean \( M = 26.1, \text{standard error of the mean } SEM = 1.9 \)) than the novices (mean \( M = 16.3, SEM = 1.5 \)). No significant interaction was found between rater and group for the GRS scores.

Analysis of the checklist scores indicated a main effect of rater, \( F(2, 26) = 5.03, p = .014 \) (Figure 3), and post hoc pairwise comparisons revealed that ratings of the hand therapist were significantly lower (mean \( M = 8.4, SEM = 5.0 \)) than those of the certified hand therapist and informed nonclinician for either measurement tool. Moorthy, Munz, Sarker, and Darzi (2003) identified one of the drawbacks to using a checklist and informed nonclinician were too lenient in critical in her marking or whether the occupational therapist and the informed nonclinician for either measurement tool. This finding is interesting because both of these evaluators were occupational therapists who had received additional training in orthosis fabrication. The finding is important because it suggests that the evaluators’ expertise may influence the scores given, particularly for the checklist, and ensuring that all evaluators have similar levels of experience may be necessary. Further studies are required to standardize the scores appropriate for levels of experience and to determine whether the certified hand therapist was overly critical in her marking or whether the occupational therapist and informed nonclinician were too lenient in their marking.

It is interesting that no significant difference was found between the scores given by the licensed occupational therapist and the informed nonclinician for either measurement tool. Moorthy, Munz, Sarker, and Darzi (2003) identified one of the drawbacks to using a checklist and GRS to assess surgical procedures as the requirement that all evaluators have similar levels of experience may be necessary. Further studies are required to standardize the scores appropriate for levels of experience and to determine whether the certified hand therapist was overly critical in her marking or whether the occupational therapist and informed nonclinician were too lenient in their marking.

**Discussion**

With the increasing workload of occupational therapy students and educators, it is important that educational programs use efficient strategies to teach and evaluate technical skills. However, to demonstrate the effectiveness of new pedagogical approaches, evaluative measurement tools are required. A measurement tool is useful to students, educators, and researchers as a measure of competence or educational effectiveness only if it measures what it claims to measure (i.e., has construct validity) and if the results are consistent across time and raters (i.e., the measure is reliable). The degree of consistency or agreement among different evaluators for a particular measurement tool is a measure of interrater reliability (Portney & Watkins, 1999; Zeller, 1990). A high level of interrater reliability is an important characteristic of a measurement tool because it ensures consistent evaluation by different evaluators, thus revealing variability in skill level rather than variability among evaluators. Both the GRS and the checklist developed for the evaluation of fabrication of MCP joint–stabilizing orthoses showed acceptable levels of interrater reliability among three evaluators with varying levels of experience with the evaluated skill. Thus, one can infer that the variability in participants’ scores is the result of variability in skill level and was only minimally affected by the different evaluators.

Although the three evaluators’ scores were moderately correlated, analyses conducted to evaluate construct validity indicated that when using the checklist, the certified hand therapist gave significantly lower scores than the licensed occupational therapist. This finding is interesting because both of these evaluators were occupational therapists who had received additional training in orthosis fabrication. The finding is important because it suggests that the evaluators’ expertise may influence the scores given, particularly for the checklist, and ensuring that all evaluators have similar levels of experience may be necessary. Further studies are required to standardize the scores appropriate for levels of experience and to determine whether the certified hand therapist was overly critical in her marking or whether the occupational therapist and informed nonclinician were too lenient in their marking.

**Table 1. Interrater Reliability of Three Raters Using a GRS and Checklist to Assess Videotaped Performances and Final Products of Orthosis Fabrication**

<table>
<thead>
<tr>
<th></th>
<th>Absolute Agreement Method</th>
<th>Consistency Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICC^a</td>
<td>95% CI^b</td>
</tr>
<tr>
<td>GRS</td>
<td>.76 (0.53, 0.90)</td>
<td>.79 (0.58, 0.92)</td>
</tr>
<tr>
<td>Checklist</td>
<td>.79 (0.59, 0.92)</td>
<td>.79 (0.58, 0.92)</td>
</tr>
<tr>
<td>GRS and Checklist</td>
<td>.88 (0.79, 0.94)</td>
<td>.88 (0.79, 0.94)</td>
</tr>
</tbody>
</table>

Note. CI = confidence interval; GRS = global rating scale; ICC = intraclass correlation coefficient.

^a Single-measure ICC based on three raters. ^b 95% CI for estimate of ICC.
an experienced surgeon must take time from surgical practice to observe and score the trainees’ performances. Recruiting a certified hand therapist or even a licensed occupational therapist for evaluation of trainees may be difficult and expensive. Recruiting licensed occupational therapists who have not necessarily specialized in hand therapy or training nonclinicians to evaluate students may help to make these tools more accessible and practical to use. Further studies examining the interrater reliability among several licensed occupational therapists, other certified hand therapists, and trained nonclinicians would be beneficial to confirm these findings.

Evidence of construct validity in surgery is generally found in the measurement tool’s ability to differentiate between novice and experienced surgeons on the performance of a given task (Vassiliou et al., 2005). This same method of validation was applied to the GRS and checklist. Both were found to be valid indicators of the skills involved in creating this specific orthosis. That is, they were able to differentiate between novices and experienced therapists by showing a significant difference in scores between these groups for each evaluator’s scores. Therefore, we believe that these tools accurately measure skill level in creating MCP joint-stabilizing orthoses.

Figure 2. Mean global rating scale scores (with standard error bars) for experienced therapists and novices, showing higher scores for experienced therapists than for novices.

Figure 3. Mean checklist scores (with standard error bars) for experienced therapists and novices, showing higher scores for experienced therapists than for novices and lower scores given by the certified hand therapist than by the licensed occupational therapist.
Global ratings and checklists are currently being used to complement each other when measuring technical skills in the field of medicine. The surgically based OSATS is one such tool, comprising a procedure-specific checklist and GRS consisting of more generic components mapped onto a Likert scale (Moorby et al., 2003; Vassiliou et al., 2005). The measurement tools developed in this study likewise consist of a procedure- and product-specific checklist and a more generic process-oriented GRS. Regehr et al. (1998), as well as Vassiliou et al. (2005), have found that checklists did not improve the reliability or validity of the measurement when both were used and suggested that the GRS is “the more appropriate marking technique and extensive task-specific checklists are not necessary” (Regehr et al., 1998, p. 996).

Moreover, we believe that GRS has at least two advantages over the checklist. First, because the GRS is not specific to a particular type of orthosis, validating its use for evaluating fabrication of other orthoses will be more convenient. Second, the nature of the GRS lends itself to providing more detailed and critical feedback to trainees who would like to know the areas in which they can focus efforts for improvement. It is interesting, however, that our results do not fully support the idea that the GRS should be used by itself because the agreement and consistency between raters (as measured by ICC values) increased when scores from both the GRS and the checklist were used (see Table 1). These differing conclusions may be the result of features of the checklists used (e.g., length), this particular technical skill, or the validation process used.

We acknowledge that although our results have shown that the GRS and checklist we developed demonstrate interrater reliability and construct validity, further studies are needed to evaluate other psychometric properties, such as test–retest reliability, intrarater reliability, and responsiveness. Earlier, we discussed validity as a property of the measurement tool. However, other researchers have suggested that a more practical and useful approach is to consider validity as a property of the application of measurement tools, that is, the testing environment (Hodges, 2003; Howley, 2004), because validity is contextual and may vary with features of the test environment, such as fidelity of the hand, communication with the client, or completeness of the task being evaluated. As such, further studies are needed to validate the tools in other contexts.

We also believe that future tests should build on these tools to develop evaluations for fabrication of other types of orthoses and to incorporate assessment of communication skills (e.g., LeBlanc et al., 2009) and evaluate novel educational methods in occupational therapy, such as simulation. Uses for evaluating simulation methods may involve evaluating the educational usefulness of both non-pathological and arthritic model hands for training in technical skills of orthosis fabrication. If simulation training proves to be a safe, efficient, and effective method of training people in the fabrication of orthoses, we hope that this educational method will be adopted by occupational therapy educators in teaching other technical skills.

In summary, we have demonstrated a method for developing a checklist and GRS for evaluating orthotic fabrication skills that demonstrate both interrater reliability and construct validity. We believe that these measurement tools will help to address the need for objective, valid, and reliable measures to evaluate students’ performance and skill level and to explore and evaluate various educational methods. Results suggest that although a certified hand therapist may be more critical than a licensed occupational therapist in evaluating students, training nonclinicians to perform these assessments may be possible, which may save time and money for occupational therapy departments. Nonetheless, the GRS and checklist are practical for evaluating students and providing them with feedback on their technical skills as well as for providing researchers and educators with information regarding the effectiveness of educational interventions.

**Implications for Occupational Therapy Practice**

The results of this study have the following implications for occupational therapy practice:

- The Delphi method can be used to develop valid and reliable measurement tools for orthosis fabrication skills in an occupational therapy curriculum.
- Validated measurement tools—a GRS and a checklist—now exist that can be used to evaluate trainees’ technical skills when fabricating MCP joint–stabilizing orthoses (contact Heather Carnahan for copies). The measurement tools can be used or adapted for research evaluating the effectiveness of training interventions for orthosis fabrication, such as simulation of pathologically deformed limbs. ▲

**Acknowledgments**

We thank Sammons Preston Canada Inc. for donating the thermoplastic material used in this study as well as Marie Eason-Klatt from St. Joseph’s Health Centre in Toronto and two anonymous Delphi panelists for their expert advice during the Delphi process. This research was funded.
by an award from the University of Toronto Educational Development Fund awarded to Heather Carnahan and Pat McKee as well as the BMO Financial Chair in Health Professions Education Research awarded to Heather Carnahan. The work described in this article was presented at the 2008 Thelma Cardwell Research Day at the University of Toronto.

References


