User interface issues with infusion pumps: A systematic review and guidelines for usability testing

Jan Maarten Schraagen¹, Ph.D., Martin Schmettow², Ph.D., & Rutger van Merkerk³, Ph.D.

¹TNO & University of Twente; ²University of Twente; ³University Medical Center Utrecht
Overview

› ‘Patient Safety’ Project

› Systematic literature review

› Issues in Human Factors Validation Testing:
  › Number of test participants required
  › Population of intended users

› Empirical test to address issues

› Conclusions and recommendations
Patient Safety Project

- Three-year project (2011-2013), aimed at improving patient safety through two main tracks:
  - Improving human-technology interaction
  - Improving human-human interaction

- Improving human-technology interaction:
  - Cooperation between academic teaching hospital, university, research organisation, small business
  - Focus on improving interfaces for infusion pumps
  - Approach: literature review, user requirements elicitation, prioritizing user requirements through interviews, iterative interface design development, usability testing
Systematic literature review\(^1\)

- Review (\(N=55\)) showed wide variety of usability issues, but also many non-usability issues (e.g., issues during pump procurement process)

- Many opportunities where things can go wrong during the infusion process, particularly considering widespread pump usage

- 76 requirements were derived from the literature, grouped in 9 use cases (e.g., placement/removal syringe; administer bolus; start/stop infusion)

- Interviews with total of 7 ‘super-users’ from 3 departments (OR, ICU, Nursing) led to prioritizing of requirements and confirmed validity

---

\(^1\) Schraagen et al. (under review). User-interface issues with infusion pumps: A systematic review. Submitted to J Biomed Informatics
From requirements to testing: Issues in Human Factors Validation Testing

› In our interviews, we noted differences in priorities among departments

**Issue #1:** Generalization regarding safe and effective use by the ultimate population of users

› Medical devices constitute high-risk equipment, for which established standards\(^1\) for estimating sample sizes in usability testing may be inappropriate

**Issue #2:** Evidence-based determination of sample size in high-risk equipment

\(^1\) e.g., Virzi, 1992; Nielsen, 1993, Lewis, 2001; Faulkner, 2003
Current quantitative control strategies for use in evaluation studies

A priori control (Nielsen):
- Set goal for detection rate D
- Set number of users to 5
- Run study

Early control (Lewis):
- Set goal for detection rate D
- Run a few sessions
- Estimate number of users
- Run study

Late control (Schmettow):
- Set goal for detection rate D
- Run session
- Estimate # of remaining defects
- Terminate study

Goal?

No

Yes
Limitations of previous approaches

- Virzi’s formula underestimates remaining number of defects, due to variance in defect visibility

- Medical devices may differ from software products in problem detection rates

- Goals for detection rate $D$ may need to be set higher for medical devices than ~80%-85%, leading to higher numbers of users to be tested than the standard five recommended by Nielsen

---

Current study\(^1\)

- Goal: to evaluate a late control strategy with a medical device for different user groups

- High-level method:
  - Develop novel interface design for infusion pump
  - Select representative user groups (OR and ICU)
  - Select representative tasks for users to carry out with infusion pump
  - Observe user problems and apply ‘triage strategy’ (sanitize dataset)
  - Apply late control strategy:
    - Set goal D = 90%, with 90% CI
    - Run session with subsample and estimate # remaining defects
    - Continue until goal D = 90% is reached

Novel interface design, presented on touch-screen
User groups

OR anaesthesiologists: N=18
ICU nurses: N=18
2 participants excluded due to incomplete video data

Level of education (distribution)

Pump experience (years)

Age (distribution)

OR

ICU

total average

20-29
30-39
40-49
50-59
Tasks

- Fixed set of 11 tasks covering main functions of infusion pump

- Typical tasks: interpreting the meaning of an alarm, adjusting values, and checking pump status

- Tasks were piloted with three experts and were assessed as being:
  - Externally valid
  - Of roughly equal difficulty
  - Independent of each other
Procedure

- Setting: controlled, quiet environment in 1,042-bed academic teaching hospital
- Consent form and Pre-task demographics questionnaire
- Think-aloud while performing 11 tasks (video and audio recorded as well as screen captures)
- After each task: immediate retrospective think aloud (1 minute)
- After all 11 tasks were performed, three post-task questionnaires:
  1. 72-item design features questionnaire (5-point Likert scale)
  2. 2-item semantic differential scale on CTA experience
  3. Exterior appearance semantic differential scales
- Full procedure completed within 90 minutes
Triage

Think-aloud

Questionnaire

Expert judgment

“Full (‘raw’) data set”

“Stripped data set”

Definitely not a usability problem

Definitely a usability problem

Undecided

Baltimore, March 12, 2012
HFE in Health Care

“Full (‘raw’) data set”

“Stripped data set”

Definitely a usability problem

Definitely not a usability problem

Undecided
Ergonomical and cognitive design principles (Rubin, 1994)
Results

Phase 1
- $N=10$ (OR)
- $N=10$ (ICU)
109 (89) problems observed

Phase 2
- $N=7$ (OR)
- $N=7$ (ICU)
86 (75) problems observed
Example progress analysis

Process figures

Progress figures (%D)

*First phase ICU-trials (N=10)
Quantitative results

<table>
<thead>
<tr>
<th>Raw data set</th>
<th>User group</th>
<th>LNB-fit</th>
<th>N</th>
<th>5Seen</th>
<th>6X=0</th>
<th>% (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1</td>
<td>OR</td>
<td>1AnPh1</td>
<td>10</td>
<td>91</td>
<td>19</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>ICU</td>
<td>NuPh1</td>
<td>10</td>
<td>83</td>
<td>34</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>OR+ICU</td>
<td>3Ph1</td>
<td>20</td>
<td>109</td>
<td>24</td>
<td>82</td>
</tr>
<tr>
<td>Phase 2</td>
<td>OR</td>
<td>AnPh2</td>
<td>7</td>
<td>69</td>
<td>81</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>ICU</td>
<td>2NuPh2</td>
<td>7</td>
<td>74</td>
<td>43</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>OR+ICU</td>
<td>Ph2</td>
<td>14</td>
<td>86</td>
<td>25</td>
<td>77</td>
</tr>
<tr>
<td>Combined (phase 3)</td>
<td>OR</td>
<td>4An</td>
<td>17</td>
<td>107</td>
<td>37</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>ICU</td>
<td>Nu</td>
<td>17</td>
<td>95</td>
<td>27</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>OR+ICU</td>
<td>All</td>
<td>34</td>
<td>123</td>
<td>31</td>
<td>80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stripped data set</th>
<th>User group</th>
<th>LNB-fit</th>
<th>N</th>
<th>5Seen</th>
<th>6X=0</th>
<th>% (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1</td>
<td>OR</td>
<td>1AnPh1</td>
<td>10</td>
<td>74</td>
<td>11</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>ICU</td>
<td>NuPh1</td>
<td>10</td>
<td>73</td>
<td>23</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>OR+ICU</td>
<td>3Ph1</td>
<td>20</td>
<td>89</td>
<td>12</td>
<td>88</td>
</tr>
<tr>
<td>Phase 2</td>
<td>OR</td>
<td>AnPh2</td>
<td>7</td>
<td>61</td>
<td>136</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>ICU</td>
<td>2NuPh2</td>
<td>7</td>
<td>64</td>
<td>18</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>OR+ICU</td>
<td>Ph2</td>
<td>14</td>
<td>75</td>
<td>20</td>
<td>79</td>
</tr>
<tr>
<td>Combined (phase 3)</td>
<td>OR</td>
<td>4An</td>
<td>17</td>
<td>87</td>
<td>20</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>ICU</td>
<td>Nu</td>
<td>17</td>
<td>80</td>
<td>12</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>OR+ICU</td>
<td>All</td>
<td>34</td>
<td>98</td>
<td>11</td>
<td>90</td>
</tr>
</tbody>
</table>

← AnPh2 D = 46%
< 85%

← AnPh2 D = 31%
> 85% → more efficiënt data set
Contribution problems detected once

Large contribution of problems detected once in the **full data set** compared to **stripped data set**
Problem distribution between anesthesiologists and ICU-nurses

Some problems are observed (more often) by either of the user groups.
Conclusions

› On the basis of the raw data set, the goal of 90% detection was never reached, not even with the combined sample of \( N=34 \)

› Goal of 90% detection was only reached when the data set was stripped of problems that were definitely not a usability problem AND when the two user groups were combined (\( N=34 \))
  › Extrapolation under these assumptions to \( D=95\% \) leads to \( N=66 \), and for \( D=98\% \) to \( N=129 \)

› Model-based predictions of problems detected are highly sensitive to:
  › Individual differences in experience levels
  › Problems mentioned only once
  › User groups
Number of users required for each major user group, for goal D=90% (stripped data set)

N=31 OR users
90% CI: .78-.95

N=21 ICU users
90% CI: .80-.97
Recommendations

› Do not use the magic number approach, use the late-control strategy instead

› Pay more attention to the quality of the data set, and use triage-like methods to sanitize the data set

› Variance in defect visibility exists and may lead to gross underestimates of the number of users required, particularly when different user groups need to be taken into account