Evaluating the Quality of Multiple-Choice Tests with Automatically Generated Visualizations

An Application of Scalable Vector Graphics in Medical Education

Dennis Toddenroth, Dr. med Thomas Frankewitsch
What does testing in medical education look like?

An exemplary one-best-answer item:

What is the cause of chickenpox?

a) The bacterium *Bordetella pertussis*
b) Morbillivirus
c) Rubella virus
d) Varicella zoster virus <- correct
e) Human herpesvirus type six (HHV-6)
What makes a good test item?

- Common assumption: The level of difficulty should not be extreme. An item will not reveal information about the relative competence of any candidate if everybody answers correct or incorrect.

- The item should truly measure 'competence'. In this case 'relatively competent' students are typically (although not necessarily) more likely to answer correctly.

- 'Competence' is presumed to be recognizable by better overall scores, so the set of remaining items is considered to allow the assessment of the presumed individual 'competence'. 
Rationale for graphics creation

Limiting the analysis to the established parameters 'p-value' (defined as the fraction of correct choices) and 'discrimination index' (the correlation with the residual scores) is not always demonstrative and sometimes neglects information such as problems with single distractors.

"We recommend that attention be focused on the pattern of responses rather than on the difficulty level or discrimination index." ¹

"Indeed graphics can be more precise and revealing than conventional statistical computations." ²


Motivation for developing a custom-made solution

- Perpetual requirement to evaluate large amounts of data: Operating an off-the-shelf statistics software could become tedious.

- Evaluation procedure quite uniform, affording a reasonable level of automation

- Fine-grained control over layout, ability to structure outputs as needed

- General enthusiasm for programming and SVG
Key steps of the evaluation process
Note: Most namespace prefixes such as „<svg:rect/>“ have been eliminated to increase readability.
Program output II

<!-- sample chart for distractor 'E)' -->
<svg width="80" height="12">
  <defs>
    <linearGradient xml:id="55647">
      <stop style="stop-color: rgb(0,160,0)" offset="0%"/>
      <stop style="stop-color: rgb(160,160,160)" offset="100%"/>
    </linearGradient>
  </defs>
  <g style="stroke:#000000; fill: url(#55647)">
    <rect style="fill:black" height="12" width="0.5" stroke-width="0.5" y="0" x="39.5"/>
    <rect height="4.75" stroke-width="0.5" width="17.12" x="22.38" y="3.375"/>
  </g>
</svg>

<table>
<thead>
<tr>
<th>in dieser Gruppe richtig</th>
<th>Δ in %-Pkt. (besser ➤)</th>
</tr>
</thead>
<tbody>
<tr>
<td>67.16% (4271 von 6359)</td>
<td></td>
</tr>
<tr>
<td>65.74% (706 von 1074)</td>
<td>-1.42</td>
</tr>
<tr>
<td>69.51% (1493 von 2148)</td>
<td>+2.35</td>
</tr>
<tr>
<td>69.10% (4599 von 6656)</td>
<td>+1.94</td>
</tr>
<tr>
<td>58.60% (1993 von 3401)</td>
<td>-8.56</td>
</tr>
</tbody>
</table>
### Program output III

<!-- sample boxplot for candidate distribution 'A)' -->
<svg width="320" height="14">
  <g style="stroke:#000000; fill:#999999">
    <!-- ... -->
    <!-- maverick: -->
    <circle r="1.5" cy="7" cx="49.599"/>
    <!-- whiskers: -->
    <line y2="10" y1="4" x2="76.759" x1="76.75999"/>
    <line y2="10" y1="4" x2="277.56" x1="277.56"/>
    <line y2="7" y1="7" x2="277.56" x1="76.75999"/>
    <!-- quartiles: -->
    <rect height="10" width="67.040" y="2" x="163.56"/>
    <!-- median: -->
    <line y2="13" y1="1" x2="192.64" x1="192.64"/>
    <!-- mean: -->
    <circle style="stroke:#000000; fill:#000000" r="1.5" cy="7" cx="190.32"/>
  </g>
</svg>

<table>
<thead>
<tr>
<th>Median (Interquartilsabstand)</th>
<th>Boxplot (besser ➤)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A)</strong> 68.16% (60.89% - 77.65%)</td>
<td>![Boxplot A]</td>
</tr>
<tr>
<td><strong>B)</strong> 69.0%</td>
<td>![Boxplot B]</td>
</tr>
<tr>
<td><strong>C)</strong> 73.47% (57.96% - 81.28%)</td>
<td>![Boxplot C]</td>
</tr>
<tr>
<td><strong>D)</strong> 68.44% (62.85% - 75.98%)</td>
<td>![Boxplot D]</td>
</tr>
<tr>
<td><strong>E)</strong> 60.34% (50.84% - 66.48%)</td>
<td>![Boxplot E]</td>
</tr>
</tbody>
</table>
Program output IV

<path style="stroke:black;stroke-width:0.02;fill:url(#rocGrad13611)"
d="M1 0 L1.0 0.0142 L1.0 0.0428 L0.975 0.0428 L0.926 0.328 ..."/>
Program output V

Observation: A variety of diagrams can be consizely composed from the set of available SVG elements.
Observations on performance

- Evaluating several test items and assembling a multi-page report with numerous graphics can take minutes.

- As a consequence this approach appears not just yet ready for on-demand computation.

- Bottlenecks seem to be data aggregation and rendering PDF (involves parsing XML and producing layout).

- Workaround: a web-based frontend where users request reports. This application version subsequently emails download links, which refer to zipped collections of individually prepared files.
Klinische Chemie

Fachsemester: 1. Klinisches Semester
Prüfungssemester: SS 2008
Anzahl Fragen: 20

Inhalt & Auswertung zu...

\[ \text{Frage 129} \]

---

Median (Interquartilsabstand) | Boxplot (besser >)
-------------------------------|-----------------------------------

<table>
<thead>
<tr>
<th>40%</th>
<th>80%</th>
<th>80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>A) 66.48% (59.78% - 74.86%)</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>B) 60.9%</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>C) 67.32% (56.84% - 73.6%)</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>D) 68.16% (60.89% - 78.21%)</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>E)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Schwierigkeitsindex: \( p = 0.64 \) / Trennscharfe: \( r = +0.24 \)
Alternative output target: XHTML+Javascript+SVG

- Firefox directly renders SVG elements integrated into XHTML.

- Javascript makes DOM-elements accessible, allowing image parts to become responsive and mutable (navigatable).

- The interactivity of the clickable scatterplot aims at making the identification of the odd one out easier.

- Most users have incompatible IE, configuration not (yet?) widely applicable
Output utilization

- PDF-reports are displayed on monitors or printed in grayscale. Vector graphics ensure a constantly impeccable visual appearance.

- Efficient PDF file sizes permit electronic report distribution without compromising image quality.

- The reports provide accessible feedback to test authors (and hopefully support the creation of high-quality test items).

- The reports are incorporated into decisions eliminating faulty items from grade evaluation.
More observations

- The design has been in operation for roughly 2½ years. The reports were gradually expanded along the way.

- Initial program versions were rather 'clipboard-composed' from FOP's examples.

- Approximately 8,500 test items and a total of 1,000,000 individual choices have been processed so far.