Investigating Effects of Common Spreadsheet Design Practices on Correctness and Maintainability

Daniel Kulesz
Institute of Software Technology
University of Stuttgart
daniel.kulesz@informatik.uni-stuttgart.de

Sebastian Zitzelsberger
SebastianZitzelsberger@gmx.de

6th July 2012
The Starting Point

- John receives an list with 350 spreadsheets who were originally developed by Jane. Jane left the company 2 weeks ago.

- Most of Jane's spreadsheets are still in use today.

- The boss asks John to “identify and sort out the risky ones” in order to “prevent future use prior careful inspections”.

“Is this a good spreadsheet?”

- Dunn [Dunn, 2010] divides spreadsheets into three „classes“:
  - the good
  - the bad
  - the downward ugly

- To help classify spreadsheets, Dunn proposes to check them for certain “spreadsheet characteristics” ( = product-oriented Best Practices).

Can spreadsheet characteristics really help in quality assessment?

[Dunn, 2010]: Dunn, A: Spreadsheets - The good, the bad and the downward ugly, Proceedings of the EuSpRIG 2010 conference
Related work

- Hardcoded constants (“jamming errors”) are significantly positively correlated with faults (“logic errors”) [Teo and Tan, 1997].

- Faults in more complex formulae have a higher chance of staying undetected after inspections [Panko, 1999].

- High formula complexity decreases correctness [Kruck, 2006].


Research Questions

- **RQ1**: Can observance of spreadsheet Best Practices serve as a dependable quality indicator?
  - **RQ1a**: Does observance of certain Best Practices affect correctness?
  - **RQ1b**: Does observance of certain Best Practices affect maintainability?

We evaluated RQ1a and RQ1b for the following Best Practices:

- Do not put constants in formulae
- Keep formula complexity low
- Reading direction (“Refer to the left and above”)
Agenda

- Definitions
- Our Experiment:
  - Setup
  - Evaluation Method
  - Observations
- Validity Evaluation and Summary
Definitions
Anomalies in Software

(according to IEEE Std 1044-2009)

- Problem (i.e. bad decision)
- Failure (i.e. wrong result)
- Error (i.e. wrong click)
- Fault (i.e. wrong reference)
- Defect (i.e. wrong reference)
- Rule Violation* (i.e. "bad" reading direction)

* our extension, not part of the IEEE Std 1044-2009

- Human

- Every fault is a defect, but not every defect is a fault.

- "blameful" or "blameless"

- May cause

- May introduce
Quality

- Quality according to DIN 55350 (translated from German):
  
  “Quality is the sum of properties and characteristics of a product or an activity which contribute to the suitability of satisfying described needs.”

- Correctness (our definition):
  
  “The degree to which the “bottom-line” results computed by a spreadsheet conform to manually computed results for given test scenarios.”

- Maintainability (our definition):
  
  “The sum of internal factors of a spreadsheet which contribute to its chance of maintaining correctness throughout modifications.”
Our Experiment: Setup
Overview

- Two Phases:
  - Phase 1: **Creation** of new spreadsheets from scratch
  - Phase 2: **Modification** of existing spreadsheets (initially created in Phase 1)
- Task descriptions (for both phases) on A4 paper sheet plus additional A4 paper sheets with data
Subjects

- Total of 42 subjects
  - Phase 1: 18 subjects
  - Phase 2: 24 subjects
- Mixed population of secretaries, engineers, students, teachers, artists and others (ages between 20 and 55)
- Subjects from Phase 1 did not participate in Phase 2 (with one exception)
Phase 1 - Task

- **Task:** Develop a comparison sheet for passenger cars ("Which product to buy?")

**Requirements:**

- must allow to compare at least 3 passenger car models by tuning weightages
- no macros allowed
- must be precise to the first decimal
- should be clearly arranged and optically appealing
Phase 1 - Data

- We provided 3 “Testimonials”.
- Testimonials rate aspects (i.e. quality of the seats) and give them ratings in the German grade system:
  - 1 = very good
  - 2 = good
  - 3 = average
  - 4 = satisfactory
  - 5 = poor
- Categories and “main categories”
- Data volume: 33 ratings for each car
Phase 1 - Sample solution

![Spreadsheet Image]
Phase 2 - Task

Task: Extend the comparison of Phase 1

Requirements:

- add two additional testimonials (provided by us).
- introduce a new variable “annual mileage” which impacts two existing categories with a handicap.
- Manipulate the parameters so that a certain result is achieved.
- Visually mark the “winner car”.

Erweiterter Fahrzeugvergleich
Ein Spreadsheet-Experiment

Hintergrund
Angesichts des vielfältigen Angebots an PKW am Markt stehen viele Neuausgangslösungen bevor der Wahl. Um die Entscheidung in solchen Fällen zu erleichtern, hat ein 'virtueller personalisierter Kollege aus Ihrem Team ein Spreadsheet (eine „Excel-Tabelle”) erstellt, das für einige erklärte Fahrzeuge eine persönliche Gesamtbewertung errechnet. Das Spreadsheet verwendet dabei zwei Dauererheben als Ausgangsbasis:

1. (Das sogenannte Testzeugnis, das in allen veröffentlichten PKW-Testen der Abteilung “Test und Technik” der ADAC-Zentrale abgedruckt und Einzelheiten für verschiedene Bewertungsaspekte (z.B. die Laufzeit des Motors) enthält.
2. (Vom Nutzer individuell festlegbare Gewichtungen für die einzelnen Bewertungsaspekte (der Nutzer kann damit ausdrücken, dass ihm beispielsweise das Kauferhalten eines Fahrzeugs wichtiger als sein Verbrauch ist.)

Aufgabe
Ihre Aufgabe besteht darin, das Spreadsheet ihres ehemaligen Kollegen an eine Reihe neuer Wünsche und Anforderungen anzupassen.

Neue Wünsche und Anforderungen
- Erweitern Sie den Fahrzeugvergleich um die beiden im Anhang enthaltenen Testzeugnisse (BMW 118d RPi und Dacia Sandero 1.6 Laufleicht).
- Führen Sie die neue Datenvariable „Jährliche Fahrleistung“ ein, die sich durch ein Handicap (z.B. einen Multiplikator) anmerkbar, die von ADAC festgelegten Kriterien im Bereich „Streckenkosten“ sowie „Versicherungs- / Rentenkosten“ auswirkt, und zwar so wie in der folgenden Tabelle dargestellt:

<table>
<thead>
<tr>
<th>Jährliche Fahrleistung</th>
<th>Handicap Streckenkosten</th>
<th>Handicap Versicherung / Rentenkosten</th>
</tr>
</thead>
<tbody>
<tr>
<td>weniger als 15.000 km</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>15.000 bis 25.000 km</td>
<td>1.9</td>
<td>1.0</td>
</tr>
<tr>
<td>25.001 bis 35.000 km</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>35.001 bis 50.000 km</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Über 50.001 km</td>
<td>2.0</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Beispiel: Mit Betriebskosten BMW 118d bei 40.000 km/Jahr 1.2 * 1.7 = 2.18
- Stellen Sie die Gewichtungen und die jährliche Fahrleistung so ein, dass der Dacia Sandero die beste und der Fiat (bis zu den schlechtesten persönlichen Gesamtwerten) bekommt, ohne die Xaven in den Testzeugnissen zu manipulieren.
- Heben Sie das Fahrzeug, dass die beste persönliche Gesamtwert erhält, durch eine andere Farbe und eine größere Schrift, visuell als Empfehlung hervor.
Phase 2 - „Input sheets“

Input sheet 1 (sheet 11 in phase 1)

Input sheet 2 (sheet 14 in phase 1)

Input sheet 3 (sheet 2 in phase 1)

Input sheet 4 (sheet 4 in phase 1)
Our Experiment: Evaluation Method
Evaluating Correctness

- Manual inspection for data errors
- Manual calculation of correct values for defined test scenarios (including corner cases etc.) with calculator

**Example Test Scenario:**

Annual mileage: 20,000 km
Weight for interior (category): 5
  Weight for Usability: 7
  Weight for Front space: 9
  Weight for Back space: 1
  Weight for Interior flexibility: 1

**Expected outputs**

- Car 1: 2.3
- Car 2: 2.7
- Car 3: 2.5
- Car 4: 3.3
- Car 5: 2.1
Measuring Correctness

- # of successful scenarios
- # of failed scenarios
- # of “partially successful” scenarios
- Small excerpt for one scenario:

<table>
<thead>
<tr>
<th>Soll-Wert</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datenfehler</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Input-Sheet</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

This was the author. (not counted in overall stats)
Evaluating Best Practice Conformity

- Automated defect checks using Spreadsheet Inspection Framework (SIF).
- SIF Features:
  - Implements checking rules for “the” three Best Practices
  - Open Source
  - Customizable Thresholds (we used two configurations)
  - Supports grouping of violations

Example report
Measuring Best Practice Conformity

- Objective metrics (number of cells, number of formulae)
- Absolute number of rule violations
- Relative number of rule violations (i.e. percentage of “complex” formulae)
- Increases / decreases of the above metrics from Phase 1 to Phase 2

Example measures for a Phase 2 spreadsheet
Our Experiment:
Observations
Lots of data!

<table>
<thead>
<tr>
<th># of data errors</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td># of cells</td>
<td>220</td>
<td>530</td>
<td>397</td>
<td>592</td>
<td>222</td>
<td>497</td>
<td>214</td>
<td>780</td>
<td>223</td>
<td>741</td>
<td>507</td>
<td>382</td>
<td>220</td>
<td>336</td>
<td>237</td>
<td>444</td>
<td>419</td>
</tr>
<tr>
<td># of formulae</td>
<td>3</td>
<td>163</td>
<td>166</td>
<td>184</td>
<td>30</td>
<td>126</td>
<td>27</td>
<td>277</td>
<td>3</td>
<td>241</td>
<td>69</td>
<td>7</td>
<td>30</td>
<td>54</td>
<td>7</td>
<td>223</td>
<td>87</td>
</tr>
<tr>
<td>Failed in # of scenarios</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>1</td>
<td>1</td>
<td>X</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>X</td>
</tr>
</tbody>
</table>

| Configuration 1 | Formula Complexity | 0 | 1 | 27 | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 3 |
|------------------|--------------------|---|---|----|----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|                  | .. relative to # of formulae | 0% | 1% | 16% | 16% | 100% | 100% | 0% | 0% | 3% |
| No Constants In  | Formulae           | 0 | 10 | 67 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
|                  | .. relative to # of formulae | 0% | 6% | 40% | 0% | 0% | 0% | 0% | 0% | 3% |
| Reading direction|                   | 0 | 27 | 27 | 27 | 27 | 27 | 0 | 198 | 1 | 4 | 27 | 54 | 0 | 0 | 47 | 0 | 0 | 0 |
|                  | .. relative to # of formulae | 0% | 17% | 16% | 100% | 90% | 100% | 0% | 21% | 0% |

| Configuration 2 | Formula Complexity | 3 | 4 | 67 | 30 | 0 | 0 | 0 | 0 | 0 | 3 | 30 | 54 | 0 | 3 | 3 | 3 |
|------------------|--------------------|---|---|----|----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|                  | .. relative to # of formulae | 100% | 2% | 40% | 16% | 7% | 0% | 0% | 43% | 100% | 100% | 0% | 1% | 3% |
| No Constants In  | Formulae           | 0 | 0 | 67 | 0 | 0 | 126 | 0 | 0 | 3 | 7 | 0 | 0 | 3 | 0 | 3 | 3 |
|                  | .. relative to # of formulae | 0% | 0% | 40% | 0% | 0% | 100% | 0% | 0% | 4% | 100% | 0% | 0% | 43% | 0% | 3% |
| Reading direction|                   | 0 | 27 | 27 | 184 | 27 | 27 | 27 | 0 | 198 | 1 | 4 | 27 | 54 | 0 | 0 | 47 | 0 |
|                  | .. relative to # of formulae | 0% | 17% | 16% | 100% | 90% | 21% | 100% | 0% | 82% | 1% | 57% | 90% | 100% | 0% | 21% | 0% | 0% |
Success and Failure Rates

Phase 1

Phase 2 (excluding authors)
Authors not better than non-authors

<table>
<thead>
<tr>
<th>Input Sheet</th>
<th>Success</th>
<th>&quot;Partial&quot;</th>
<th>Fail</th>
<th>Unuseable</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1 (+1)</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2 (+1)</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1 (+1)</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5 (+1)</td>
<td>1</td>
</tr>
</tbody>
</table>
# Formula Complexity vs. Correctness

<table>
<thead>
<tr>
<th># of failed scenarios</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>X</th>
<th>X</th>
<th>1</th>
<th>1</th>
<th>X</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rel. formula complexity</td>
<td>0%</td>
<td>1%</td>
<td>16%</td>
<td>16%</td>
<td>0%</td>
<td>81%</td>
<td>100%</td>
<td>1%</td>
<td>0%</td>
<td>11%</td>
<td>0%</td>
<td>43%</td>
<td>100%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Phase 1 Formula complexity

- **Ranking 1:**
  - Unuseable = filtered
  - Wrong result = 1
  - Success = 0
  - SCC* = -0.4836187

- **Ranking 2:**
  - Unuseable = 1
  - Wrong result = 1
  - Success = 0
  - SCC* = -0.4061209

* SCC = Spearman Correlation Coefficient
“Low Violation Spreadsheets” are unlikely to produce correct results.

(only successful Spreadsheets from both Phases shown below)

Relative number of defects
- 0-5%
- 6-94%
- >95%

<table>
<thead>
<tr>
<th>Category</th>
<th>0-5%</th>
<th>6-94%</th>
<th>&gt;95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex formulae</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Formulae with constants</td>
<td>5</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Formulae with “wrong” reading direction</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Rule Violations and Maintainability

- Tendency: low violation rates lead to better maintainability
  (at least if we disregard Input Sheet 1)

<table>
<thead>
<tr>
<th>Input Sheet</th>
<th>Success</th>
<th>Partial</th>
<th>Fail</th>
<th>Unusable</th>
<th>Constants in Formulae (Violations)</th>
<th>Formula Complexity (Violations)</th>
<th>Reading Direction (Violations)</th>
<th>Sheets with Data Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1 (+1)</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>16%</td>
<td>40%</td>
<td>16%</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2 (+1)</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1%</td>
<td>6%</td>
<td>17%</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1 (+1)</td>
<td>2</td>
<td>100%</td>
<td>100%</td>
<td>90%</td>
<td>1</td>
</tr>
</tbody>
</table>

What is wrong with Input Sheet 1?
Input Sheet 1 - the special guy?

- The only Input Sheet with multiple worksheets
- The only spreadsheet with “plausibility checks“
- Maintained by “advanced” Excel users:

How do you rate your experience regarding Excel?

- Input01
- Input02
- Input03
- Input04

Novice, Beginner, Advanced, Expert
Validity Evaluation and Summary
Validity - Assurance Attempts

- Mostly domain-free task (difficulty = getting it done in a Spreadsheet)
- Involves “non-input data” (unlike other experiments)
- Subjects received no training before experiments
- Controlled environmental variables like time limits, location, user settings and equipment (at least in Phase 2)
- Tried several threshold calibrations for the rules
- Tried to balance subjects in Phase 2
- Questionnaire at the end of each experiment
Validity - Threats

- Small sample size
- Pure black-box approach, no defect examination (unknown false positive rates)
- Small selection of Best Practices evaluated
- Used spreadsheet samples from laboratory experiments
- All subjects from Stuttgart area
- Evaluation tool not a mature product
- Final comparison realized in a spreadsheet
Summary

- RQ: Is Best Practice conformity (BPc) a dependable quality indicator for correctness and maintainability?

- We evaluated 42 spreadsheets for BPc of three rules ("No constants in formulae", "Keep formula complexity low" and "Refer to the left and above")

- Results:
  - BPc not a dependable indicator for correctness
  - Tendency of higher maintenance success for spreadsheets with higher BPc

- Secondary contributions:
  - Unique selection of comparable spreadsheets
  - Spreadsheet Inspection Framework
Questions or comments?

Spreadsheet Inspection Framework is available here:

https://github.com/SZitzelsberger/Spreadsheet-Inspection-Framework