# Bericht zum Thema: innovative Voranalysen von Fragebögen

László Pitlik, 2011.VII.

## Einleitung

Fragebögen werden meistens scheinbar sorgfältig entworfen und alle eingetroffenen Daten im Gültigkeitsintervall der jeweiligen Frage quasi ohne weiteres ausgewertet, besonders was die Anzahl der Antworten und irgendwelche Häufigkeiten angeht. Leider diese Annäherung ist nicht sorgfältig genug. Auch dann nicht, wenn einige Kontrollfragen unter den zahlreichen Fragen versteckt sind. Die Motivationen der Befragten können sehr unterschiedlich sein: ich persönlich gebe (zu) immer randomisierte, jedoch gültige Antworten… (In diesem Bericht geht es im Weiteren leider am wenigsten darum, wie „unethisch“ sich solche Befragten verhalten mögen, vielmehr darum, was sollte man dagegen unternehmen, wenn diese Verhaltensweise direkt kaum beinflussbar ist und auch unbewusst auftreten kann: vgl. unlogisches Denken, was man ja mit Kontrollfragen an sich testen möchte…)

## Problemstellungen

Falls sich der Autor eines Fragebogens nicht darauf einstellt, dass Antworten aus diversen Gründen unlogische sein können, dann bleibt nichts Weiteres übrig, wie alles in die berechneten Ergebnisse hineinzuinterpretieren. Das Interpretationsvermögen lässt sich trainieren, falls man vor dem wirklichen Einsetzen des Fragebogens Antworten generiert (egal wie), und diese versucht auszuwerten. Diese vorherige Auswertungsübung kann unter anderem darauf hinweisen, dass man nicht einmal genau weiß, was man durch Antworten erhofft. Diese Technik hilft also zwar fundiertere Fragen zu stellen, aber lässt immer noch nicht die Bösen Jungs (wie ich es bin) dabei virtuell ertappen, als sie höflich aber quasi randomisierte Antworten liefern wollen…

## Lösungsansatz

Falls sich man an die Ähnlichkeitsanalyse wendet, dann darf man erwarten, dass die Befragten und die Fragen selbst anhand einer Risikoskala bewerten lassen. Die Ähnlichkeitsanalyse erlaubt nämlich ein OAM (Objekt-Attribut-Matrix) in mehreren Zyklen auszuwerten. Die Objekte sind hierbei die Befragten, die Fragen sind die Attribute. Die Antworten sind die zu analysierende Werte. Nur solche Fragen können bearbeitet werden, wo eine Art ideal Antwort aus der Sicht der Befragung existiert. Nehmen wir ein Beispiel: Möchte man gern wissen, wie zufrieden die Befragten mit einem Informationssystem sind, dann erwartet man letztendlich die Antworten, die Zufriedenheit ausstrahlen. Hat man die Rangfolge der Antwortoptionen pro Frage, hat man ein OAM mit ranking values. Fehlende Antorten der Befragten werden mit der größten ranking value bestraft, d.h. mit der niedrigsten Chance zum Gut werden versehen: und hier beginnt die tatsächliche Arbeit…

Im Bericht wurde als Beispiel eine wirklich durchgeführte Befragung als Datengrundlage genommen. Unter den mehr als 500 auswertbar eingestuften Antwortserien gab es 75 Objekte, bei denen das Geschlecht, die Häufigkeit und Ort, wie und wo sie das geprüfte System nutzen das Gleiche war. (Ein Ergebnis im Voraus angesprochen: Wenn man alle Befragten gleichzeitig bewertet wurden, wo die diversesten Untergruppen existierten, ergab sich sofort, dass die Antworten quasi als randomisiert angesehen werden sollten, da sich keine der potentiellen Modelle anpassen ließ).

### Die Analyseschritte

Bei der mehrschichtigen Analyse es wird zunächst eine Gesamtbewertung durchgeführt (vgl. BSC = balanced score card oder COCO Y0-Modell). Hierbei wird untersucht, ob solche Transformationen (vgl. Treppenfunktionen) existieren, welche uns erlauben, alle Befragten als gleichgut an dem Ideal gemessen einzustufen. Diese Analyseschicht liefert bereits Stoffe für weitere spezielle Risiko-Interpretationen (z.B. welche Befragten verhielten sich an den anderen Befragten gemessen mehr negative als positiv).

In der Fallstudie ergaben sich folgende Ergebnisse:



Abbildung 1. Objekte (OID) und Güteunterschiede (Quelle: eigene Berechnung)

Die Abbildung 1 stellt dar, dass sich zwischen dem Schlechtesten (links) und dem Besten (rechts) -40…67= 107 Gütepunkte Entfernung entdecken lässt. Die Anzahl der Guten (über dem Null-Punk an der Güteskala /delta/) 32 (von 75) und die Anzahl der Bösen 43 (von 75) sind. Zwischen der durchschnittlichen Bewertung der Guten und der Bösen gibt es 44 Punkte Entfernung. Die Verteilung der Gütepunkte ist nicht symmetrisch: die Guten haben anzahlmäßig weniger, aber größere Wirkungen und insgesamt sind die Guten mehr entfernt von dem neutral Null-Punkt als die Bösen…

Konsequenz: angenommen, dass sich alle Befragten gleich glaubwürdig erweisen werden, kann man behaupten, dass die Anzahl der Systemtreuen/Loyalen weniger ist, aber ihr Enthusiasmus stärker ist als die Kraftfelder der Systemgegner.

Bereits bei dieser Analysestufe könnte man eine Selbstkontrollschicht einbauen, indem die gleiche Gesamtberechnung aus einem inversen OAM gestartet wird, um prüfen zu können, wie weit das Verhalten der Kraftfelder der relative Guten und der relative Bösen symmetrisch ist. Pro Objekt kann man dabei feststellen, ob man ruhig ein Urteil aussprechen kann, oder sollte man lieber noch nicht zu dem Objekt sagen.

Die Kontrollanalyse hatte allein bei OID=231 ein Warnsignal, bei dem Objekt, der ziemlich direkt an der neutralen Zone liegt, was grundsätzlich nicht besonderes ist, aber über diesen Objekt darf noch kein Urteil fallen.

Der nächste Schritt ist alle Objekte anhand alle Fragen gegenseitig zu prüfen. Es wird dabei danach gesucht, welche Objekte liefern die am meisten unlogischen Antworten in Anbetracht der eigenen Antworten auf andere Fragen und verglichen mit allen Objektantworten. Diese Relativierung erlaubt letztlich Vermutungen darüber zu äußern, wer gab die unbrauchbarsten Antworten und welche Fragen war es, die am meisten missverstanden worden sind oder wo das Antworten am problematischsten war.

In der Fallstudie wurden als nächste Etappe gleich 33 parallele Modelle gebaut, da eben 33 Fragen im untersuchten Fragebogen vorliegen (d.h. 33 Attitude-Fragen, bei denen Zufriedenheitsoptionen von 1 bis 6 gegeben worden sind. Die fehlenden Antworten wurden mit 7 kodiert.

Nun die Fragen und deren Codes, damit die Tabellen im späteren einfacher zu verstehen sind:

x1 Mir ist bekannt, wo ich Fachartikel finden kann.

x2 Mit der Bedienfreundlichkeit des Systems bin ich zufrieden.

x3 Mit der Übersichtlichkeit des Systems bin ich zufrieden.

x4 Mit der Suchfunktion des Sytems bin ich zufrieden.

x5 Mit den vorhandenen Hilfefunktionen des Systems bin ich zufrieden.

x6 Mit den Einstellungsmöglichkeiten (meine Einstellungen, meine Favoriten) bin ich zufrieden.

x7 Mit den aktuellen internen Informationen bin ich zufrieden.

x8 Mit den aktuellen externen Neuigkeiten bin ich zufrieden.

x9 Mit den Inhalten des E-Mail Newsletters bin ich zufrieden.

x10 Mit den Einstellungsmöglichkeiten des E-Mail Newsletters (Rubriken) bin ich zufrieden.

x11 Das System unterstützt mich bei meiner Arbeit.

x12 Mit dem bereitgestellten Fachwissen bin ich zufrieden.

x13 Mit dem Fachwissen zum Themengebiet "Steuern" bin ich zufrieden.

x14 Mit dem Fachwissen zum Themengebiet "Rechnungswesen" bin ich zufrieden.

x15 Mit dem Fachwissen zum Themengebiet "Betriebswirtschaft" bin ich zufrieden.

x16 Mit dem Fachwissen zum Themengebiet "Recht" bin ich zufrieden.

x17 Mit dem Fachwissen zum Themengebiet "Wirtschaftsprüfung" bin ich zufrieden.

x18 Mit dem Fachwissen zum Themengebiet "Verwaltung" bin ich zufrieden.

x19 Mit dem Fachwissen zu IT-Themen bin ich zufrieden.

x20 Die unternehmensinternen Nachrichten sind informativ.

x21 Mit den Rubriken der Startseite bin ich zufrieden.

x22 Mit der Übersicht zu Terminankündigungen bin ich zufrieden.

x23 Mit der Übersicht zu Schulungsangeboten bin ich zufrieden.

x24 Mit den Adressdaten und Telefonlisten bin ich zufrieden.

x25 Mit dem Angebot an Formularen und Vorlagen bin ich zufrieden.

x26 Mit dem Angebot an Arbeitshilfen zum Download bin ich zufrieden.

x27 Mit dem Angebot an Stammdaten bin ich zufrieden.

x28 Mit der Online-Anmeldung zu Terminen und Schulungen bin ich zufrieden.

x29 Mit der Shop-Funktion im Bestellwesen bin ich zufrieden.

x30 Mit der Funktion zur Anmeldung zu Veranstaltungen bin ich zufrieden.

x31 Mit der Übernahme von Terminen in meinen Outlook-Kalender bin ich zufrieden.

x32 Mit der Übernahme von Kontaktdaten bin ich zufrieden.

x33 Ein Forum zum Austausch von Fachwissen halte ich für sinnvoll.

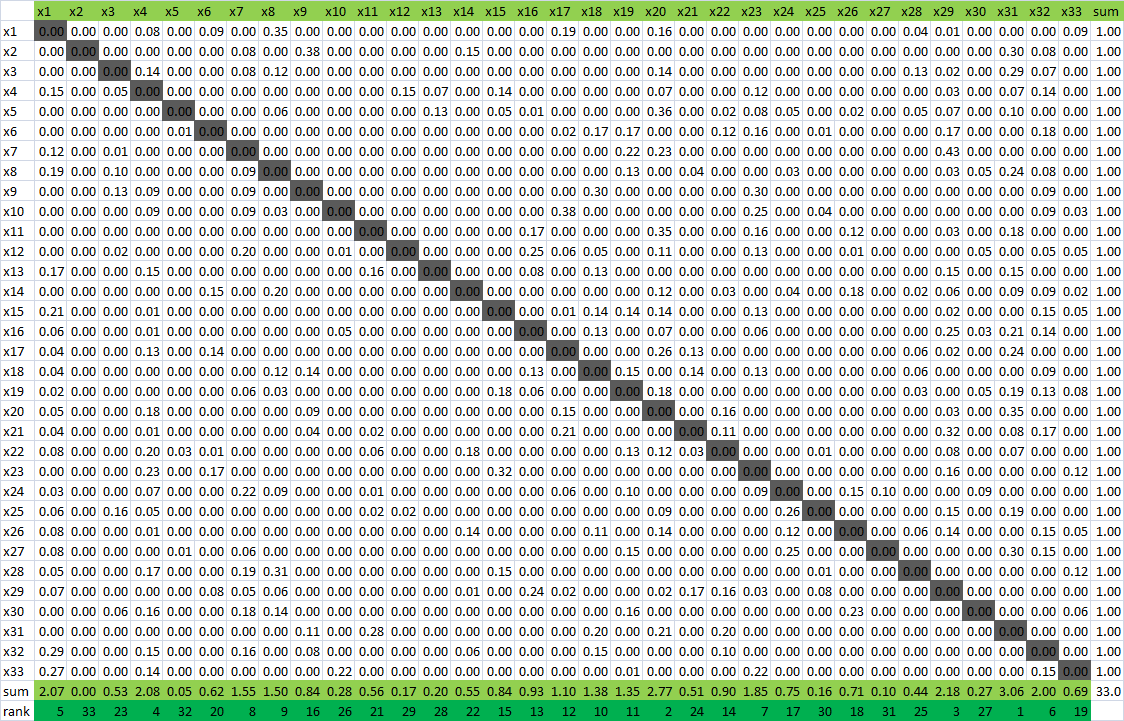


Abbildung 2: Erklärungskapazitäten der Fragen unter sich (Quelle: eigene Berechnungen)

Die Abbildung 2 stellt dar pro Zeile, wie viele Erklärungskraft der einzelnen Fragen pro Modell (Zeilen) und insgesamt aufweisen, wenn man die Antworten pro Objekt auf eine der Fragen anhand von allen übriggebliebenen Fragen schätzen möchte – angenommen, dass alle Modelle mit der idealisierten Richtungsstruktur gebaut wurden, um die Parallelität der Komponente des Ideal zu prüfen. Die Wirkungskraft wird anhand der Teilsummen pro Frage abgeleitet, wobei die Teilwirkungen pro Modell (Zeile) immer 1.00 (=100%) ergeben sollten.

Konsequenzen: Die Auswertung der Fragen die keine andere Frage an sich beeinflussen können, d.h. quasi beliebig zu beantworten sind, ist nicht zu empfehlen (z.B. x2,x5,x27,x25,...)



Abbildung 3: Erklärbarkeit von Fragen und Erklärungskraft von Fragen in Rangzahlen ausgedrückt (Quellen: eigene Berechnungen)

Die Modellfehler wurden als das prozentuale Verhältnis zwischen Schätzungsfehler und der IST-Wert berechnet entweder als absolute Zahl oder als quadratische Version. Die absoluten und quadratischen Fehler wurden als Rangzahlen abgebildet und die Summe der Fehlerversionen wurde auch dargestellt. Die Wirkungspalte ist identisch mit der letzten Zeile in der Abbildung 2.

Wie man in der Abbildung 3 deutlich sehen kann (graue Markierungen), es gibt gewisse Fragen, die weder erklärbar noch erklärend genug sind in Anbetracht aller Daten im OAM (vgl. x13,x25,x27). In diese Fragen dürfte man nicht allzu viel hineininterpretieren.

Es gibt spezielle Konstellationen auch: x20,x32,x29 ist kaum modellierbar, jedoch sie haben eine rel. starke Erklärungspotential. Außerdem x7 ist sowohl leicht zu erklären als auch rel. strak am Erklären.

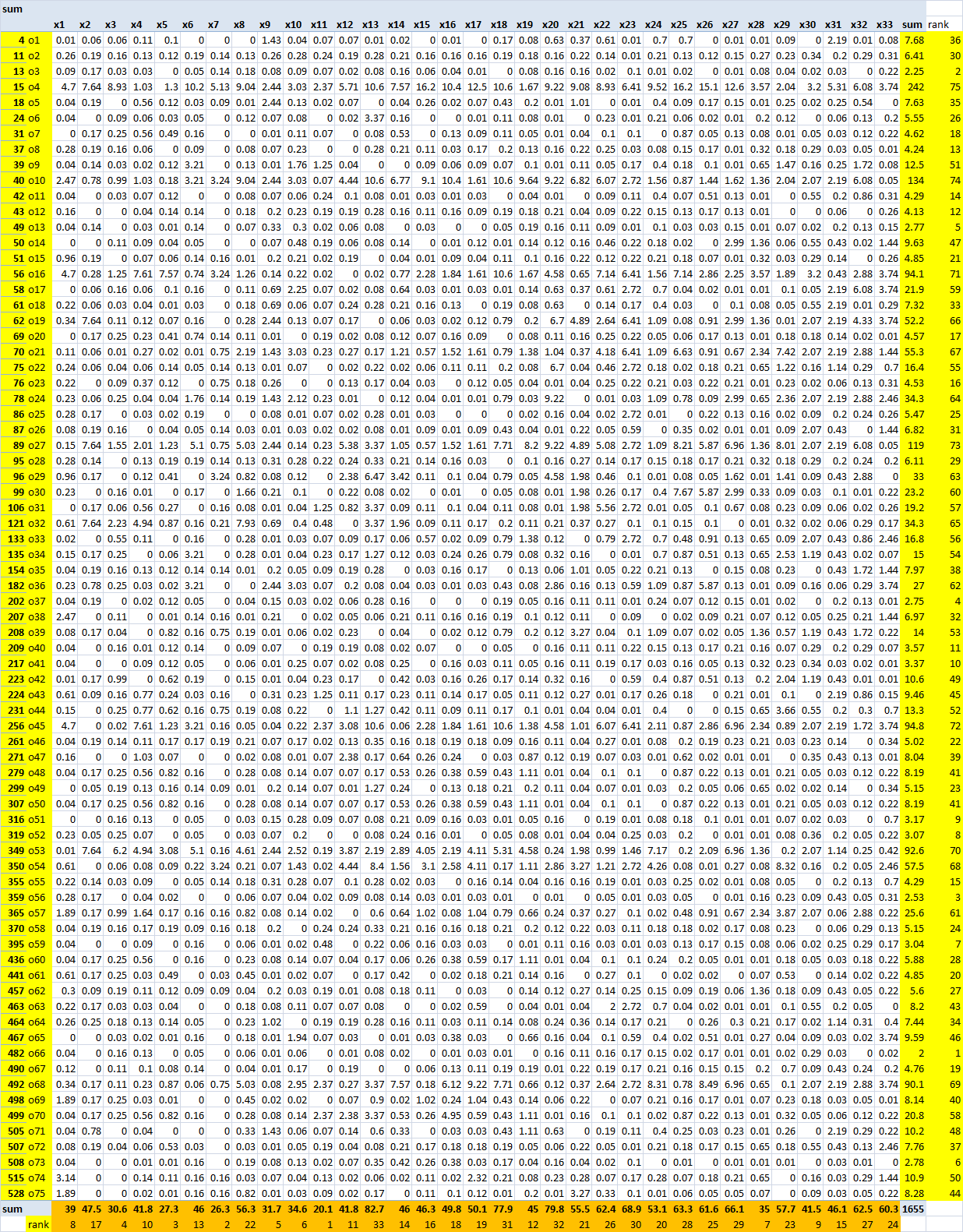


Abbildung 4: Absolute Modellfehler der Fragen (Spalten) und der Befragten (Zeilen) – Quelle: eigene Berechnungen

Abbildung 4 zeigt in der letzten Zeile die gleichen Rangzahlen, die bereits in der Abbildung 3 dargestellt wurden. Außerdem sollte man die beiden letzten Spalten an der rechten Seite näher betrachten: hier sieht man, welche Befragten am meisten unglaubwürdig sind.

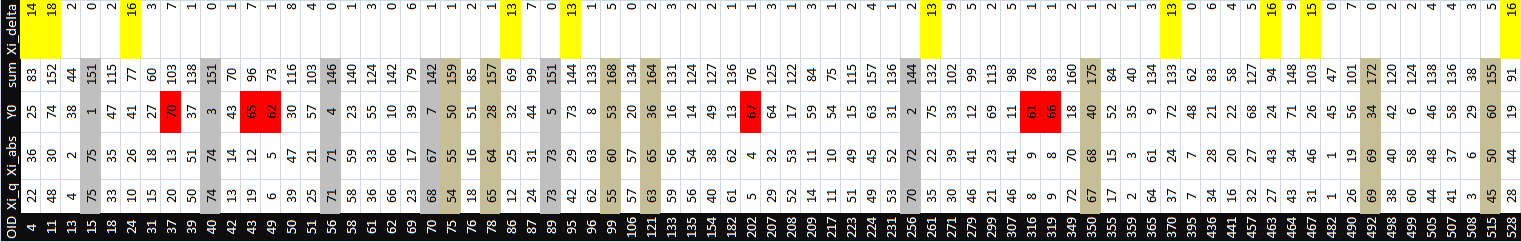


Abbildung 5: Unglaubwürdigkeit und Gesamtbewertung der Objekte (Quelle: eigene Berechnungen)

Abbildung 5 zeigt in der Zeile OID: die Identifikationscodes der Objekte. Zeile Xi wurde von der Abbildung 4 und 6 übernommen (jeweils die rechte Spalte als absolute und quadratische Objekt-Fehler). Zeile Y0 wurde aus der Abbildung 1 abgeleitet. Die Summe der beiden Rangzahlen kann man in der Zeile sum finden. Die grau-markierten Spalten zeigen die Objekte die vollkommen unglaubwürdige Antworten geliefert haben, und die sind sofort ertappt und als Streber zu bezeichnen. Die Antwortkombinationen dieser Objekte ist zwar idealistisch, jedoch aus dem Verhalten der anderen nicht ableitbar. Die andere Markierung hebt solche Objekte hervor, die scheinbar an der bösen Seite oder irgendwo in der Mitte stehen, aber dieses Verhalten ist nicht besonders überzeugend. Mit roter Markierungen sind auch Objekte zu finden, welche als glaubwürdigen Bösen einzustufen sind, da sie schlechte Idealwerte/Gesamtbewertung haben (>60), aber in beiden Objektfehlerkategorien stabile Werte haben (<20).

Abbildung 6 ist parallel mit Abbildung 4 zu interpretieren: Quadratische und absolute Modellfehler laufen zwar grundsätzlich parallel, aber es können wesentliche Unterschiede entstehen. In der Abbildung 5, die gelb markierten Spalten zeigen die Modelle, die anhand nur von einer Modellfehler-Berechnung teilweise anders betrachtet werden könnten. Diese Art von Risiko ist auch in der Gesamt Bewertung einer Befragung mit einzubeziehen.

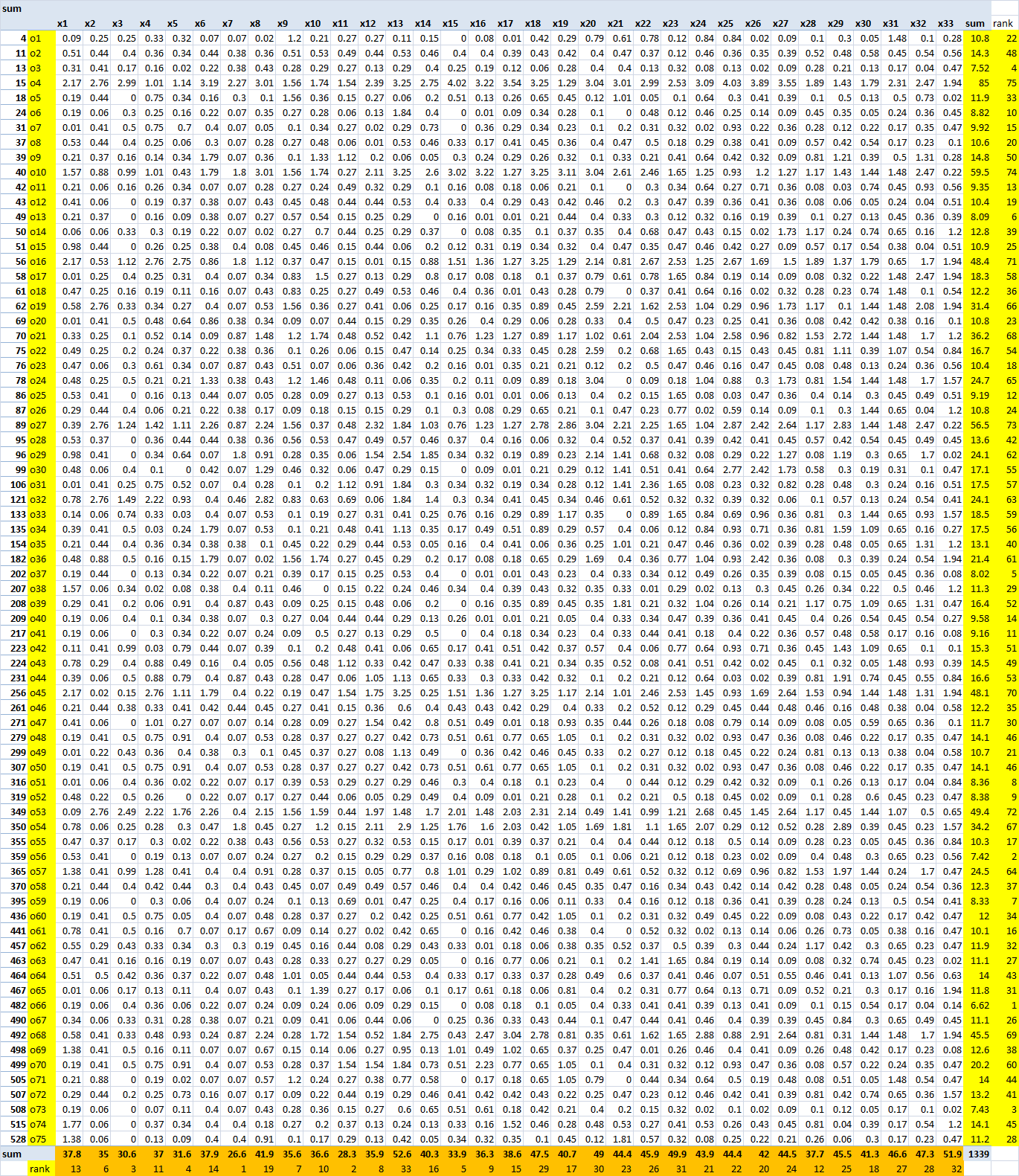
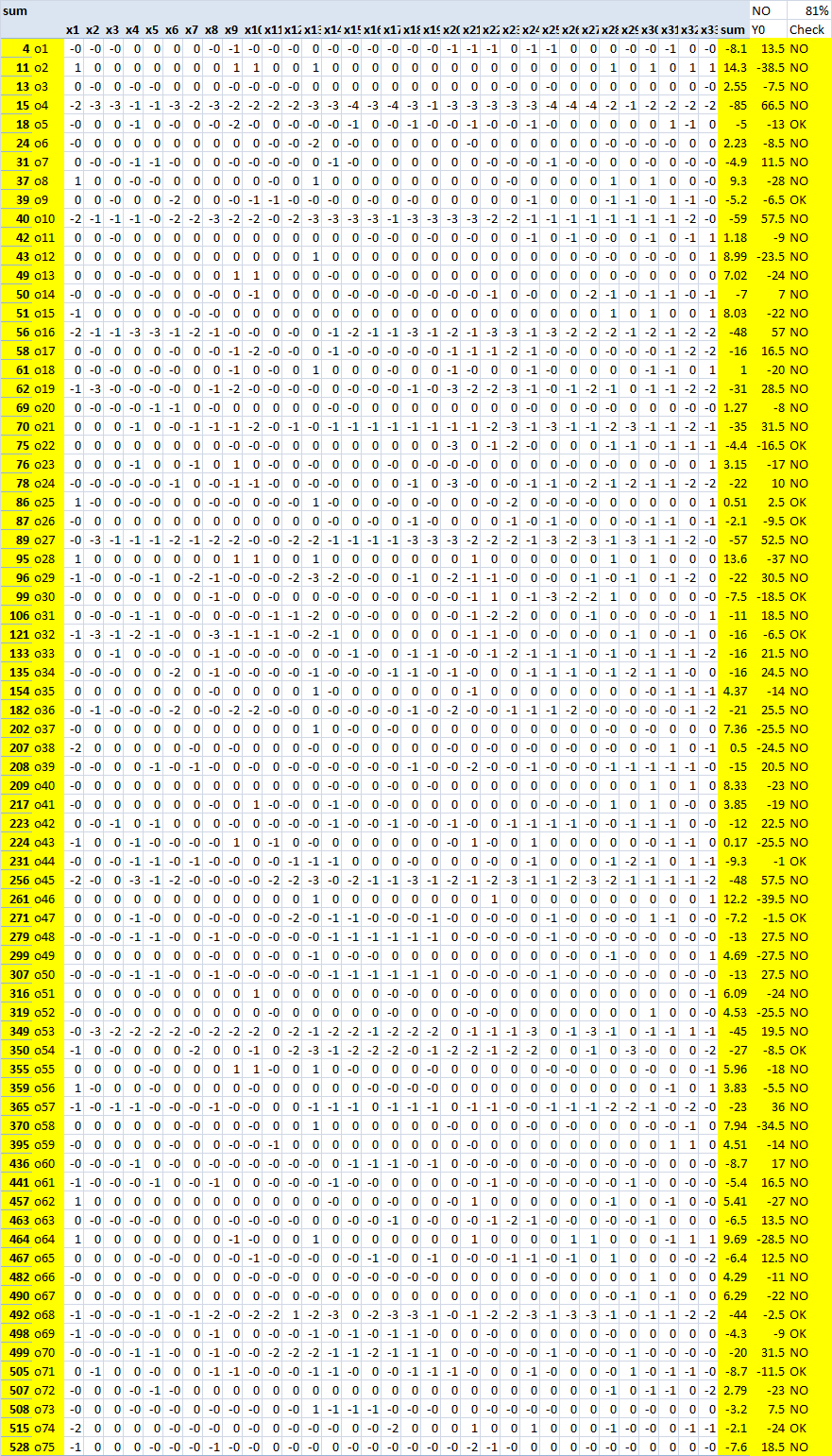


Abbildung 6: Quadratische Modellfehler der Fragen (Spalten) und der Befragten (Zeilen) – Quelle: eigene Berechnungen

Abbildung 7 leitet ein neuer Risikofaktor ab: die Parallelitätsverletzungen der Teilmodelle. Oben rechts kann man ein Prozentsatz (81%) identifizieren, d.h. nur 19 % der Objekte haben insgesamt passende Vorzeichen dem Y0-Modell gegenüber. Die Objekte mit NO Signal sind also wiederum riskanter als die anderen, da sie sich nicht gleichzeitig in allen Fragen gleich-logisch verhalten. Wo einer der gelben Werte (rechts) in den NO-Zeilen niedrig sind, dort deutet ein NO-Signal weniger auf echte Probleme als auf eine Grenzlage…

Abbildung 7: Parallelitäten der Xi und der Y0 Modelle (Quelle: eigene Berechnungen)



**Summa summarum: die 75 Befragten und die 33 Fragen sind nicht so homogen zu bewerten, als man das in einem Standardverfahren tun würde. Die gleiche Vorgehensweise lässt sich bei anderen Gruppen ebenso durchführen, sogar bis zum Detail autiomatisieren…**

## Phase II. – das Operationalisieren

László Pitlik (2011. VIII.)

Nachdem die ersten Kostproben interpretiert wurden (s. Phase I. vorab), ergab sich die Aufgabe, eine weitere Teilmenge aus den auswertbaren Fragebögen zu bilden und zwar nach Ausbildungsgruppen. Die erste Gruppe besteht aus 93 Personen, die als Wirtschaftsprüfer/in, Steuerberater/in, Steuerbevollmächtigte/r qualifiziert worden sind. Außerdem wurden Inverse-Berechnungen systematisch durchgeführt, d.h. nicht nur für das Y0 Modell sondern auch für die weiteren 33 Standard-Ähnlichkeitsanalysen.

Die erste Phase wollte den Begriff Risiko ohne konkreten/operativen Konturen darstellen, damit die Leser die noch notwendigen Interpretationen in sich durchführen. In dieser Phase werden die Risiken konkretisiert, und damit parallele Ergebnisse abgeleitet. Die bisherigen Auswertungen der Fragebögen haben z.B. einen Durchschnitt der Antworten für jede Frage ohne weiteres berechnet. Die parallelen Ergebnissen weisen auf Fragen und Personen hin, die aus diesen primären Auswertungen begründet auszuschließen sind – mit der Argumentation: die Antworten lassen eine Art Zufallsmuster vermuten!



Abbildung 8: Primäre Durchschnitte pro Frage (Quelle: eigene Berechnungen)

Die Abbildung 8. ergibt die Verteilungen der Antworten der 93 Personen. Die Bildung der gewichteten Durchschnitte wurde ohne Code 7 durchgeführt, damit die fehlenden Antworten (Code=7), die numerisch nichts bedeuten, die weiteren Berechnungen nicht unnötig beeinflussen.

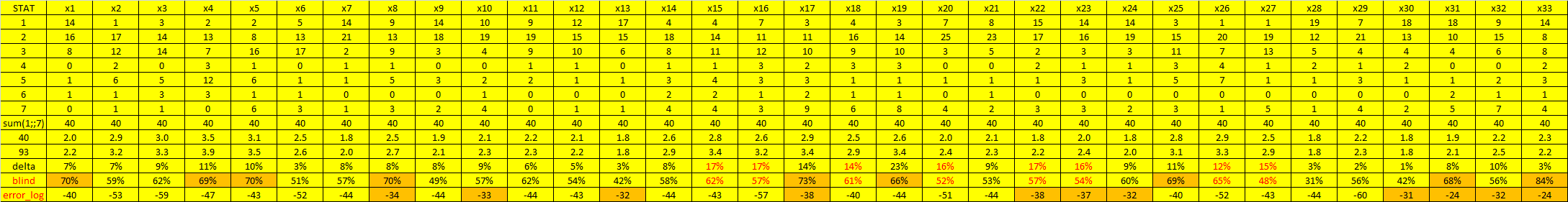


Abbildung 9a: Vergleichende Ergebnisse (Quelle: eigene Berechnungen)

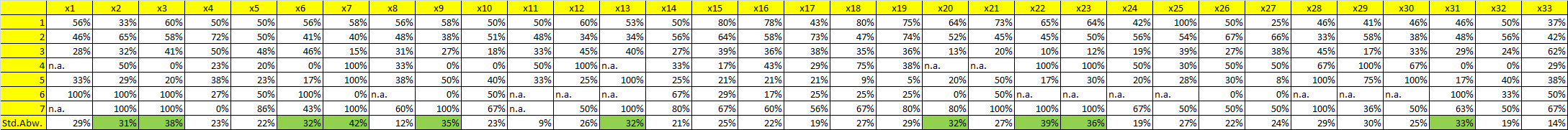


Abbildung 9b: Standard-Abweichungen anhand von Antwortverteilungen von 93 vs. 40 Objekten (Quelle: eigene Berechnungen)

Im Hintergrund der vergleichenden Ergebnisse wurde also eine ähnliche Vorgehensweise durchgeführt, wie in der Phase I: es wurden 2 Y0-Modelle gebaut (ein Direktes und ein Inverses). Diese zwei Modelle weisen darauf hin, dass sich die 93 Personen zwischen 77 Gütepunkten und 60 Fehlerpunkten verteilen (s. Abbildung 10).



Abbildung 10: Objekt ID und Y0-Punkte (Quelle: eigene Berechnungen)

Das inverse Bild hat auf keine Probleme hingewiesen, d.h. es gibt keine blinden Flecken in der Oberfläche. Diese Aussage konnte anhand der weiteren Simulationsmodellen (s. später) mit einer Sicherheit von 85% wiederspiegelt werden, d.h. 14 Objekte konnte zwischen dem direkten Y0-Modell und der Summe der 33 Teilmodelle keine passende Logik liefern.

Nach den Y0 Modellen wurden also 33 Simulationsmodelle erstellt und ebenso 33 inverse Modelle zur Kontrolle der Simulierbarkeit. Die Simulationen hatten das Ziel anhand von den jeweiligen 32 Fragen, die Antworten auf die 33. Frage zu verstehen. Die Anzahl der blinden Flecken pro Simulationsmodell ist aus der Abbildung 9 (Zeile: blind) abzulesen. Wie zu sehen ist: bis auf 5 Modelle (9;13;27;28;30) haben die weiteren Interpretationsversuche eine Risikoquote nicht selten weit über 50% (Maximum = 84%).

Die Zeile „error\_log“ in der Abbildung 9a zeigt den Logarithmus anhand von der Multiplikation der relativen Abweichungen zwischen den 93 Fakten und deren Schätzwerte pro Modell. Je grösser ein Error\_Log-Wert ist, umso unsicherer konnten die Antworten geschätzt werden. Die delta-Zeile wurde anhand der primären Ergebnisse (93 Objekte) und anhand der filterten Objekten (40 Objekte) abgeleitet. Die 40 übriggebliebenen Objekte wurde so bestimmt, dass eine minimale blind-Quote für die 14 riskanten Objekte (60,6%) festgelegt wurde und alle Objekte, dessen blind-Quote höher als diese strenge Risikogrenze liegt, wurde selektiert. Parallel ist es zu bemerken, dass die Korrelation zwischen den Y0-Werten pro Objekt und zwischen der Summe der Teilmodelle bei 0.89 lag. Dementsprechend eine Selektion der Objekte auf Glaubwürdigkeit anhand nur des risikofreien Y0-Modells scheint an sich bereits ziemlich stabil zu sein ohne 2\*33 Modelle zu bauen und deren blind-Quoten zu berechnen. Diese Selektion sollte man von der Mitte (delta=0) heraus in Richtung der extremen Werten schieben (vgl. Empfindlichkeitsanalyse). Die Summe der Teilmodelle und die Y0-Werte hatten Wiedersprüche nur im Bereich der Fehlerpunkte (0.4---22.9). Die strengere Selektion hat alle weitere Objekte als verdächtig betrachtet, wo sich die Modellierbarkeit (blind-Quoten) überdurchschnittlich ergaben. Eine ähnliche Vorgehensweise wäre gewesen, wenn die untypischen Y0-Werte zur Selektionsgrundlage gewählt worden wären.

Die Zellen mit Orange-Hintergrundfarben weisen auf Fragen hin, die lieber nicht weiter interpretiert werden sollte. Die roten Zahlen zeigen, wo extrem große Abweichungen zwischen den zwei Mengen (93 vs. 40 Objekten) vorzufinden waren.

Abbildung 9b. weist mit grün markierten Zellen auf die Fragen hin, wo die Abweichungen in der Antwortstruktur am stärksten waren. Wo die orangen und grünen Markierungen gleichzeitig zu sehen sind, dort sollte man keine Interpretationen liefern. Die Fragen die wesentliche Änderungen in der Antwortstruktur und/oder Antwortdurchschnitten aufweisen und als Interpretierbar eingestuft worden sind, können als Grundlage von eventuellen Entscheidungen dienen. Hierbei sollte streng darauf hingewiesen werden, dass die richtige Interpretation der Fragebögen immer voraussetzt, dass die Auswertungsregeln im Voraus vorliegen. Die nachträglich kreierten Bewertungen sind typische Beispiele des Hineininterpretierens und der Selbsttäuschung.

Abbildung 11 zeigt Objekt für Objekte die Selektionsgrundlagen. Abbildung 12 leitet die blind-Quoten der Objekte ab: diejenigen die hohen blind-Quoten haben, sind am meisten verdächtig Scheinantworten abgegeben zu haben.

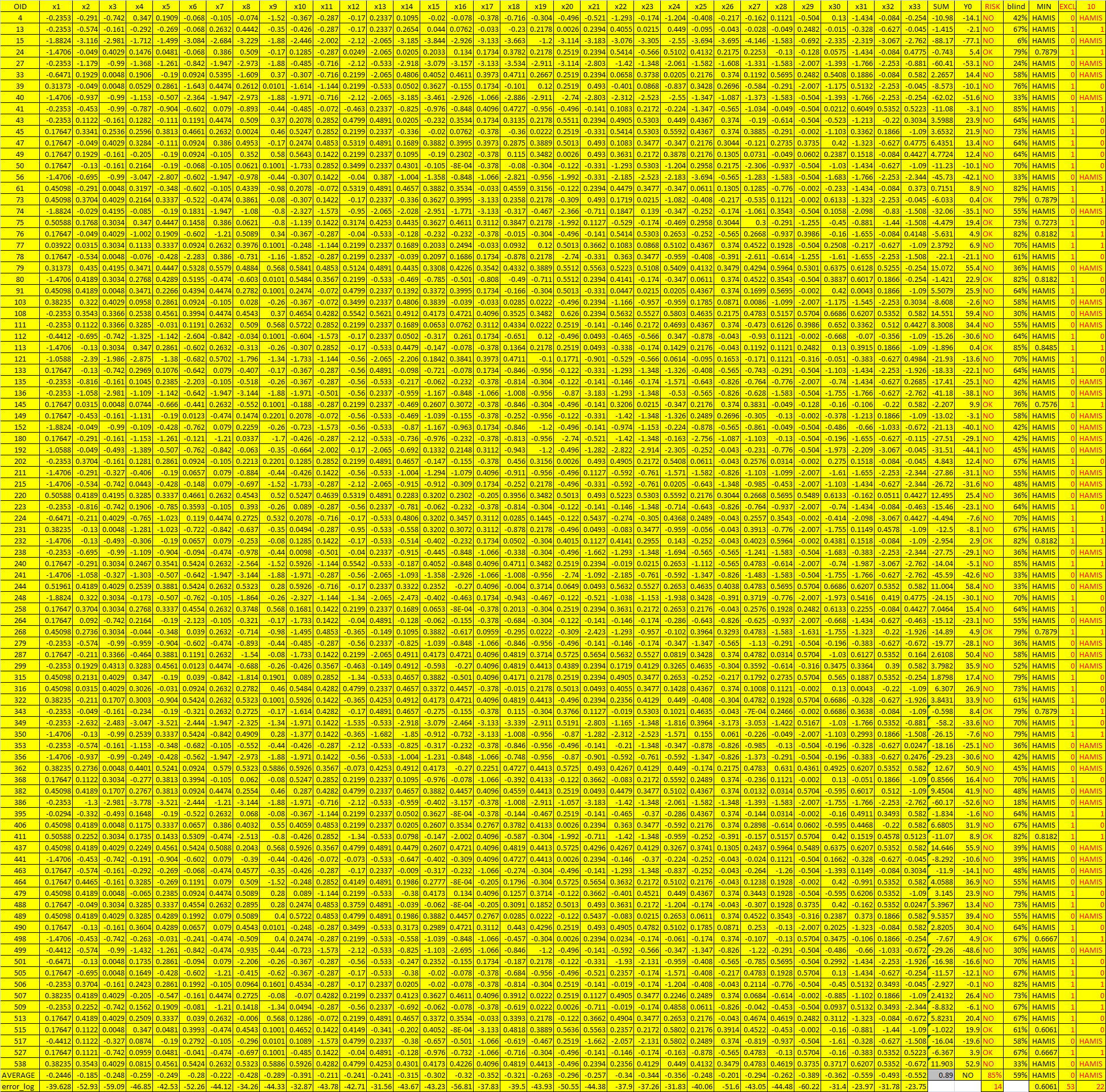


Abbildung 11: Hintergrundinformationen (Quelle: eigene Berechnungen)

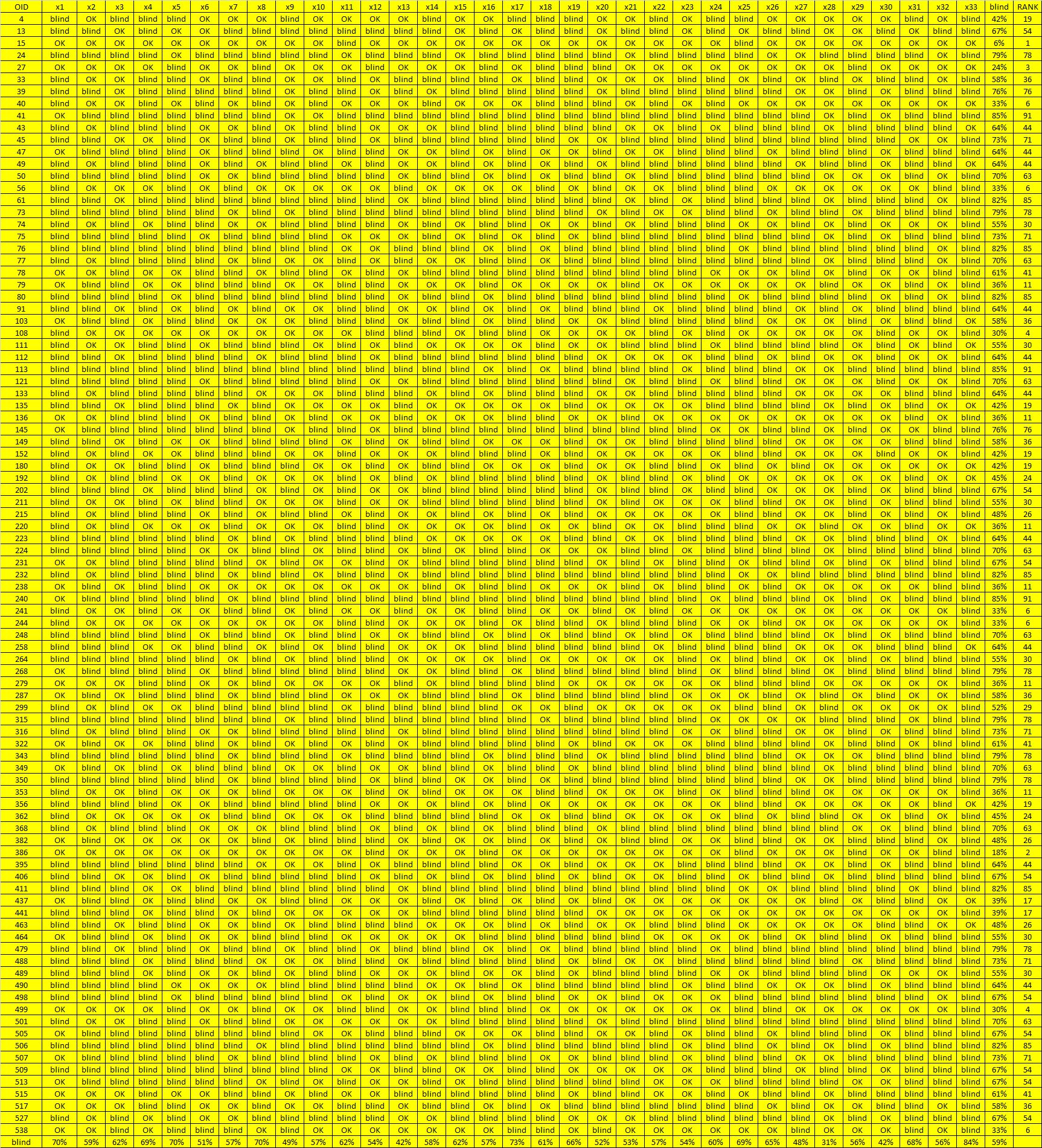


Abbildung 12: Hintergrundinformationen II. (Quelle: eigene Berechnungen)

**Summa summarum: Die 93 wichtigen Mitarbeiter (angenommen, dass nur sie befragt wurden), lieferten insgesamt ein sehr unsicheres Bild (vgl. hohe blind-Quoten pro Frage). Die Befragung darf trotzdem bewertet werden, da die Teilmodelle und das Y0-Modell insgesamt relative hohe Stabilität aufdeckten. Die gefilterten Antworten der 40 „Konsequentesten“ unterscheiden sich von der Gesamtpopulation (93 Personen) sowohl bezüglich der Durchschnitte (vgl. meistens die Übriggebliebenen lassen ein besseres Gesamtbild ausmahlen) als auch im Bereich der Antwortverteilungen in bestimmten Fällen wesentlich. Die gefilterten Personen (53) erklären die hohen blind-Quoten, d.h. die inkonsequent angekreuzten Antworten. Abbildung 8 wies auf die Fragen hin, die besser nicht interpretiert werden sollten. Über die Selektionsprinzipien in der Voranalyse muss man ebenso im Voraus entscheiden! Die Abweichungen dürften erst durch die im Voraus festgelegten Interpretationsregeln wirken…**

## Phase III. – Kontextuelles Interpretieren

In diesem Kapitel sollten die bislang anonym behandelten Fragen inhaltlich geprüft werden, um entdecken zu können, welche kontextuelle potentielle Zusammenhänge zu identifizieren sind…