



DoE-Happen #62: Warum Faktoreffekte oft unsichtbar bleiben – Teil 4: Wechselwirkungen, die alles überlagern Teil 5/17

"Wir haben den Versuch gemacht – aber wir sehen da keinen Effekt."

Diese Rückmeldung höre ich in meinen Workshops leider häufiger.

Es sind Situationen, in denen die Erwartung an eine saubere Ursache-Wirkung-Beziehung plötzlich ins Leere läuft: Drei Faktoren wurden sorgfältig geplant, linear modelliert, im besten DoE-Gewissen analysiert – **aber: „kein Haupteffekt signifikant. Nichts. Null“**

Wo liegt der Denkfehler?

Frustr macht sich breit – denn eigentlich war man sich sicher, dass zum Beispiel der pH-Wert einen Einfluss haben müsste. Schließlich wurden genau diese Rahmenbedingungen und Einflussgrößen in wi(e)derkehrenden Taskforce-Meetings intensiv diskutiert und evaluiert.

In genau solchen Fällen kommt fast immer dieselbe Anschlussfrage:

„Liegt das jetzt an der Modellierung? Oder war der Faktor wirklich nicht relevant?“

Natürlich kann das mehrere Ursachen haben, von denen ich hier nur einige kurz nennen möchte, die erweiterte Liste finden Sie bei Bedarf im Blog 58 ([Link](#))

- Die Faktoren wurden **einzeln und nacheinander** untersucht z.B. mit OFAT
- Die Faktoren wurden **nicht im richtigen Bereich** untersucht und zeigen deshalb keine Effekte.
- **Einstell- oder Messbarkeits-Probleme** überlagern die Versuche
- **Störgrößen** treiben Ihr Unwesen
- ...

Hier in diesem Blog möchte ich jedoch besonders auf einen Punkt eingehen, der in der Praxis gerne stillschweigend vorausgesetzt wird:

Die Annahme, dass sich ein Zusammenhang rein linear – also ausschließlich über die Einzeleffekte der Faktoren (pH-Wert, Rührgeschwindigkeit, Temperatur) – beschreiben lässt.

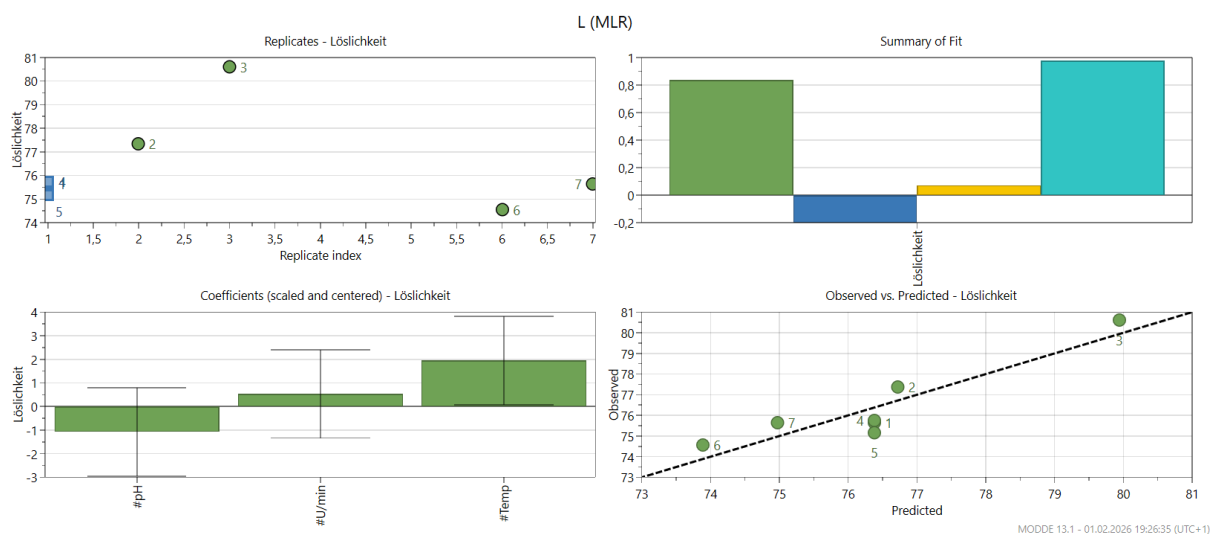
Wir schauen uns Schritt für Schritt an, warum genau diese Annahme in unserem Fall nicht trägt – und warum das in der Analyse fast zwangsläufig zu einem Modell führt, das zunächst „schwach“ wirkt oder zumindest nicht das liefert, was man sich erhofft.

Wichtig vorab: **Der folgende Datensatz / das Experiment ist schematisch konstruiert.** Nicht, weil die Realität so „ordentlich“ wäre – sondern weil ich genau *dieses* typische Praxisphänomen sichtbar machen wollte, das mir in Kundenprojekten immer wieder begegnet:

Lineare Effekte sind durchaus vorhanden – aber die Musik spielt in der Wechselwirkung.

Der Ausgangspunkt ist dabei eine sehr nachvollziehbare Entscheidung: Ein Kunde wollte aus Zeit- und Kostengründen möglichst effizient durch die Versuchsplanung kommen – in der Hoffnung, mit wenig Aufwand erste Einsparpotenziale zu heben. Da der Zusammenhang zuvor im Workshop bereits intensiv diskutiert worden war und die Variationsbreite der Faktoren als überschaubar erschien, entschied man sich für einen pragmatischen Start: **ein teilfraktioniertes Design (Resolution III) plus drei Centerpoints – also 4 Eckexperimente + 3 CP**, ausgewertet mit einem **rein linearen Modell (MLR)**.

Soweit – so pragmatisch.

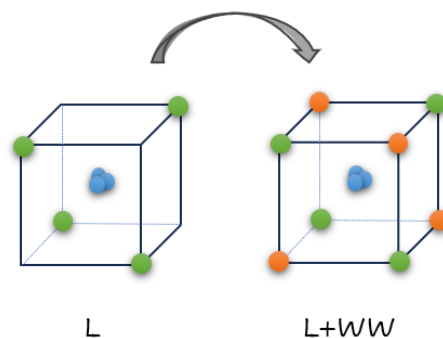


Nur leider: **In der Vorhersage war das Modell unterirdisch.** Das zeigt sich deutlich an der Vorhersagegüte (Q^2) im oberen Plot – sie ist schlicht miserabel. Die restlichen Plots erkläre ich etwas später weiter unten in der Gegenüberstellung.

Und das ist genau der Punkt, an dem viele Teams ins Grübeln kommen:
 Im reinen Fit sieht das Ganze oft noch „irgendwie okay“ aus – aber sobald man die Vorhersagegüte betrachtet (bei MODDE typischerweise über Q^2 mit Leave-one-out Cross Validation), fällt das Modell in sich zusammen.

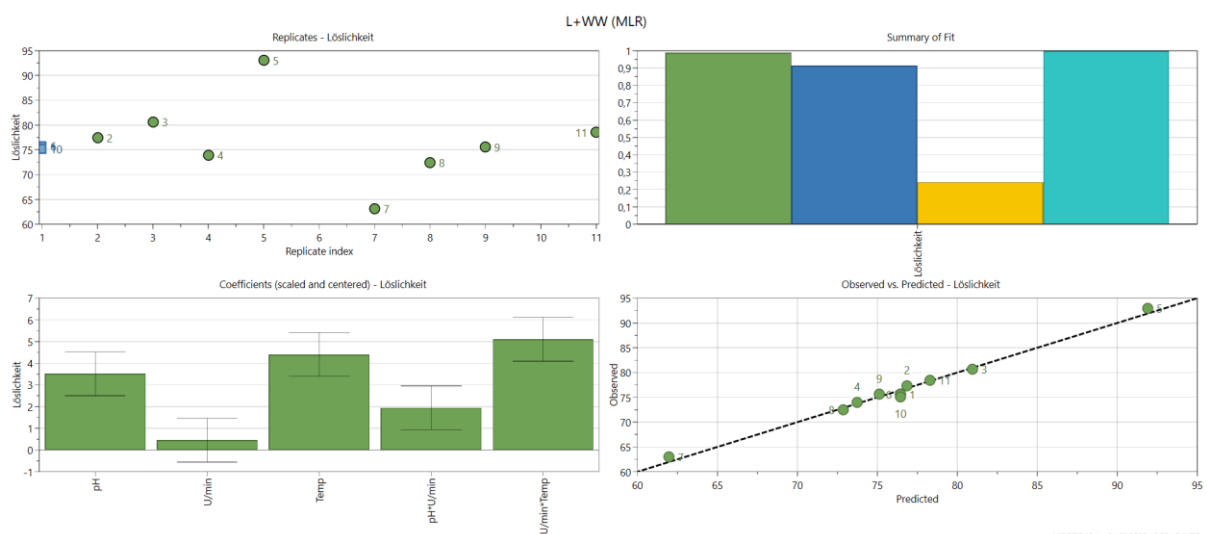
Heißt: **Es erklärt zwar scheinbar etwas – aber es sagt schlecht voraus.** Und damit wird es für Entscheidungen gefährlich.

Ein Teilnehmer, der die Versuche gefahren hatte, versicherte mir dabei sehr überzeugend: „*Ich bin sauber vorgegangen – ich glaube an die Reproduzierbarkeit.*“ Daran hatte ich auch keinen Zweifel.



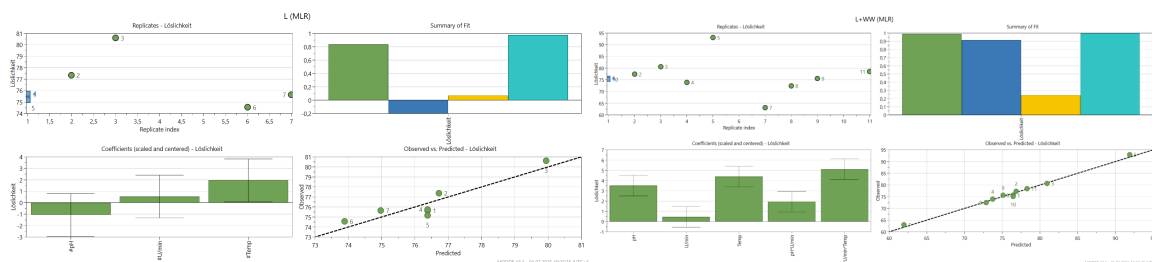
Nur war ergänzendes Experimentieren zunächst „politisch“ und budgetseitig nicht vorgesehen. Um keine voreiligen Schlüsse in den Raum zu stellen (die am Ende nur Motivation kosten), bekam der Mitarbeiter den Auftrag, die fehlenden Eckpunkte nachzulegen – also das Design per Fold-over zum Vollfaktorplan zu ergänzen, um mögliche unterschätzte Wechselwirkungen sichtbar zu machen.

Und genau hier wurde es spannend:



Nach dem Ergänzen zeigte sich, dass **nicht nur die linearen Effekte eine Rolle spielen**, sondern dass **mindestens eine Wechselwirkung wirklich treibend ist**. Dies wird schnell ersichtlich mit einem Q^2 (blauer Balken) deutlich über 0,9 – ein klarer Hinweis darauf, dass das Modell nun sehr gut vorhersagt. Auch wird klar, dass insbesondere die Kombination aus **Rührgeschwindigkeit × Temperatur**. Zusätzlich kann (je nach Datensatz/Projektlogik) auch eine zweite Wechselwirkung wie **pH-Wert × Rührgeschwindigkeit** relevant erscheinen.

→ Wichtig dabei: **Rührgeschwindigkeit „alleine“ wirkt im linearen Modell gerne unauffällig** – aber man kann sie nicht einfach rauswerfen, wenn an ihr eine relevante Wechselwirkung hängt. Denn sonst entfernt man aus dem Modell genau den Term, der den Zusammenhang überhaupt erst erfassbar macht.



Was zeigen die Plots – und worauf lohnt sich der Blick?

Im Folgenden sieht man sehr schön den Unterschied zwischen

- **Links: L (MLR) = rein lineares Modell (links)**
und
- **Rechts: L + WW (MLR) = lineares Modell + Wechselwirkung(en)**

Ich gehe die vier Standard-Plots kurz durch – so, dass man sie beim nächsten Kundenprojekt direkt wiedererkennt:

1) „Residual/Replicate“-Plot (oben links)

Dieser Plot hilft uns, die **Stabilität der Messung und die Streuung** einzuschätzen – also: *Wie sauber ist mein Experiment „handwerklich“ gelaufen? (Auch wenn die wenigen Punkte hier nicht aussagekräftig sind, so hilft der Plot doch sehr bei umfangreicheren Versuchsplanungen)*

Wenn die Punkte über die Replikate bzw. den Index wild springen oder einzelne Ausreißer auffallen, ist das ein Hinweis auf **Störgrößen, Drift oder Messprobleme**.

In diesem Beispiel sieht man: Die Streuung ist da – aber nicht so, dass man hier ohne Ursachenforschung nicht weiter machen könnte. Das ist wichtig, weil es die Interpretation stützt:

Das Problem ist nicht primär die Reproduzierbarkeit – sondern die Modellform.

2) Summary of Fit (oben rechts)

Das ist meist der Plot, an dem in Meetings zuerst hängen geblieben wird – zurecht.

Typischerweise zeigt er vier Kennzahlen (je nach Darstellung):

- **R²**: Wie gut beschreibt das Modell die vorhandenen Daten (Fit)?
- **Q²**: Wie gut sagt das Modell neue Daten voraus (Cross Validation)?
- **Model Validity**: ein übergeordneter Check, ob die Modellform grundsätzlich zum Datenverhalten passt.
- **Reproducibility**: Verhältnis zwischen Mess-/Replikatstreuung und Gesamtvariation

Und genau hier sieht man in der Gegenüberstellung das typische Muster:

Im rein linearen Modell (links) ist **Q² schwach** (oder zumindest deutlich schwächer als gewünscht) – im Modell **mit Wechselwirkung(en)** (rechts) wird Q² plötzlich stark.

Das ist der zentrale Aha-Moment: Nicht „DoE hat nichts gefunden“, sondern: „Wir haben mit einem zu einfachen Modell gerechnet – und damit die entscheidende Dynamik übersehen.“

3) Koeffizienten-Plot (unten links)

Hier sieht man die **Effektstärken der Terme** (skaliert/zentriert), inklusive Unsicherheit (Konfidenzintervalle).

Einfach gesprochen:

- **Balken groß** = Term wirkt stark
- **Konfidenzintervall schneidet Null nicht** = statistisch „stabil“ / signifikant

Der Plot zeigt sehr schön, warum das lineare Modell in die Irre führen kann:

Die **Einzeleffekte** (pH-Wert, Rührgeschwindigkeit, Temperatur) können durchaus „irgendwie signifikant“ aussehen – aber der wirklich große Hebel entsteht erst, wenn die passende 2-Fach **Wechselwirkung** mit drin ist (z. B. **Rührgeschwindigkeit × Temperatur**).

Und die anderen Wechselwirkungen (> 2-Fach z.B. $a*b*c$)? Die werden oft *mitgeschätzt*, sind aber nicht zwingend relevant – und liefern manchmal eher Rauschen als Erkenntnisgewinn.

4) Observed vs Predicted (unten rechts)

Das ist der „Reality Check“ für die Modellgüte.

- Perfekt wäre: alle Punkte liegen auf der Diagonalen.
- Je stärker die Punkte streuen oder systematisch abweichen, desto eher beschreibt das Modell die Daten **nicht** gut.
- Beim rein linearen Modell sieht man häufiger Abweichungen / Systematik.
- Mit Wechselwirkung(en) kleben die Punkte deutlich enger an der Diagonalen.

Heißt: Jetzt ist das Modell nicht nur „schön gerechnet“, sondern wirklich prädiktiv, da wir auch zwischen den Beobachtungen sicher interpolieren können.

3.) Reflexion

Wenn ich auf solche Projekte zurückblicke, dann ist es selten „die Statistik“, die uns das Leben schwer macht – es ist eher unsere intuitive Erwartungshaltung. Wir *wollen* glauben, dass sich „komplizierte“ oder umgangssprachlich „komplexe“ Prozesse sauber in Einzeleffekte zerlegen lassen: *pH hoch = gut, Temperatur hoch = besser, Rührgeschwindigkeit runter = schlechter* ... so in der Art.

Und genau hier liegt die Falle: Ein Faktor ist nicht „gut“ oder „schlecht“ – er ist oft nur im richtigen Zusammenspiel „wirksam“.

Genau deshalb werde ich häufig hinzugerufen: um die Situation aus einer ergänzenden, außenstehenden Perspektive zu beleuchten. Ich bin dabei weniger durch die „eingespielte“ Fachexpertise im Team gebunden, sondern kann durch gezielte Fragen Ziel, Ergebnis und Vorgehensweise bewusst hinterfragen – ohne dass das als Kritik gemeint ist.

Welche Rolle ich dabei einnehme, ergibt sich oft erst aus den Gesprächen und der Aufarbeitung: mal der DoE- oder DMAIC-Experte, mal eher derjenige, der methodisch strukturiert, moderiert und das Team wieder fokussiert – der Facilitator, der hilft, dass man nicht im Klein-Klein stecken bleibt. Und ja: manchmal geht es auch darum, Reibungswärme aus dem System zu nehmen – mit einer Prise Moderation.

Denn seien wir ehrlich: Wer DoE (noch) nicht im Werkzeugkasten hat, begegnet dem Thema häufig zunächst mit den klassischen Begleiterscheinungen jedes Change-Prozesses: organisierter Widerstand, langsames Lernen und schnelles Vergessen. Wenn man das nicht begleitet, bleibt auch das, was Tuckman als Team-Entwicklung beschreibt, gerne auf der Strecke – und das Projekt verliert Geschwindigkeit, bevor es überhaupt richtig Fahrt aufnimmt.

In unserem Fall hat das teilfraktionierte Design übrigens nicht „versagt“. Es hat exakt das geliefert, was es liefern kann: ein schnelles Screening – unter der Annahme, dass die Welt halbwegs linear ist und Wechselwirkungen nicht dominieren. Genau diese Erkenntnis hilft uns, aus dem Tunnel auszubrechen und zu erkennen: Sobald eine Wechselwirkung (hier: Rührgeschwindigkeit \times Temperatur) stärker ist als die Einzeleffekte, wird es heikel.

Denn dann kann ein lineares Modell zwar Effekte anzeigen – aber trotzdem nicht prädiktiv sein. Und genau das sieht man am Q^2 .

Das ist auch der Grund, warum ich in Workshops so gerne (leicht provokant) frage:

„Woran würdet ihr merken, dass ihr gerade ein Modell optimiert – das die Realität gar nicht beschreibt?“

Dieser kleine Perspektivwechsel hilft enorm. Denn das passiert schneller als man denkt: Man diskutiert über Balken im „Koeffizientenplot“, feilt am p-Wert – und übersieht dabei, dass die Punkte im „Observed vs Predicted Plot“ eben nicht dort liegen, wo sie liegen sollten. Und

im Alltag zählt genau das – etwa wenn man die optimale Einstellung sucht, nicht nur eine nachträgliche Erklärung.

Zusammenfassung

Viele „unsichtbare“ Faktoreffekte sind in Wahrheit keine fehlenden Effekte – sondern überlagerte Effekte. Häufig steckt eine dominante Wechselwirkung dahinter, die in einem rein linearen Ansatz nicht (oder nur verzerrt) sichtbar wird.

Teilfraktionierte Designs (Res III) sind effizient – aber sie kaufen diese Effizienz mit Annahmen. Wenn Wechselwirkungen wichtig sind, können lineare Modelle im Frac-Fac „blass“ wirken oder in der Vorhersage abstürzen (Q^2 niedrig), obwohl einzelne lineare Terme scheinbar signifikant sind.

Centerpoints helfen, aber sie „heilen“ nicht die falsche Modellform. Sie stabilisieren die Schätzung (und manchmal wird Q^2 besser) – aber wenn die Hauptmusik in der Wechselwirkung spielt, bleibt das lineare Bild unvollständig. Und wenn zusätzlich Nicht-Linearitäten ins Spiel kommen (wie in Blog #61), helfen Centerpoints vor allem dabei, genau diese Krümmung überhaupt erst zu erkennen bzw. abzuschätzen.

In unserem Beispiel kam der Aha-Moment erst mit der Design-Ergänzung (Fold-over / Full-Factorial). Sobald die relevante Wechselwirkung Temperatur \times Rührgeschwindigkeit modelliert wird, wird das Modell prädiktiv (Q^2 hoch) – und die Interpretation wird plötzlich stimmig.

Praxisregel für robuste Modelle: In Mehrfaktordesigns lohnt es sich fast immer, 2-fach-Wechselwirkungen bewusst mitzudenken – während 3-fach- und höhere Wechselwirkungen in der Praxis häufig mehr „Rauschen“ als Erkenntnis bringen (und gleichzeitig den Versuchsaufwand massiv in die Höhe treiben).

Und ein kleines Highlight zum Schluss

Alle bisherigen DoE-Beiträge gibt's jetzt gebündelt in unserer neuen **LinkedIn-Gruppe**. Dort finden Sie ältere Inhalte schnell wieder, können neue Themen anregen – und mit anderen DoE-Interessierten in den Austausch gehen.

👉 Alle 1–2 Wochen kommt ein frischer Blogpost dazu.

Ich freue mich auf Ihre Fragen, Erfahrungen und Diskussionen!

Hier geht's zur Gruppe:

<https://lnkd.in/d8t4gt74>

Und natürlich: Alle Blogs wie gewohnt auch auf meiner Website – inkl. Videos, Tipps & mehr.



Mehr aus Ihren Prozessen herausholen?

Ob DoE-Einstieg oder knifflige Spezialthemen wie Screening, Robustheit, Mischungen oder Troubleshooting – ich begleite Sie praxisnah:
mit Trainings, Beratung und methodischer Unterstützung – vom ersten Workshop bis zur Umsetzung.

Auch bei MVDA, DFSS oder QFD stehe ich gern an Ihrer Seite.

📧 Fragen? Schreiben Sie mir direkt: info@stefan-moser.com

🌐 Mehr unter: www.stefan-moser.com

🔗 Zur DoE-Community auf LinkedIn: <https://lnkd.in/d8t4gt74>