A Paradigm for Developing Better Measures of Marketing Constructs

Gilbert A. Churchill (1979), S. 64-73
In: Journal of Marketing Research, 16 (): 64-73

ABSTRACT

1) Nach Jacoby (1978) sind viele Instrumente der Marketing-Forschung schlecht und dienen dem Prinzip ‚garbage in - garbage out'. Diese sollten aber den wichtigsten Kriterien zur Evaluierung von Instrumenten - Validität, Reliabilität und Sensitivität – genügen. Unter Rückgriff auf die Psychologie-Forschung entwickelt der Autor ein Framework zur Evaluierung von Marketing-Instrumenten.

2) Jede Messung (X_m) besteht aus einem ‚wahren' Wert (X_t), einem systematischen Fehlerwert (X_s) und einem Zufallsfehler (X_r). Ursachen dafür sind stabile Charakterzüge und transiente Persönlichkeitszustände, situative und mechanische Faktoren sowie die Fragenauswahl, der Frageninhalt und die Erhebungsart. Als valide (Methode: face, convergent, discriminant) gilt, wenn ein Instrument misst was es soll (X_0 = X_t) und als reliabel (Methode: split-halt, test-retest), wenn unabhängige aber vergleichbare Instrumente dasselbe Ergebnis liefern (X_r = 0). Reliabilität / Konsistenz ist eine Notwendige Bedingung für Validität. Das folgende Verfahren (Pkt. 3-6) gilt nur für Multi-Item Skalen, die besser sind als Single-Item Skalen (zu spezifisch, zu kleine Gruppen im Ergebnis, Fragereihenfolge nicht prüfbar aber hat Einfluss).

3) Die Definition des Konstruktes sollte explizit bezeichnen, was in die Variable gehört und was nicht. Dies ist durch eine Literaturrecherche mit einem Statement über die Notwendigkeit eines neuen / anderen Konstruktes zu versehen.

4) Literaturrecherche, Expertengespräche und Produktvergleichsstudien liefern hinweise über die Dimensionen und Aspekte von jeweiligen Konstrukten. Zudem können Scenario-Studien mit Konsumten als erste Runde dienen zu einfache / schwere, ja/nein-Fragen oder Zweideutigkeiten zu verändern.

5) Reliabilitätsüberprüfung (Konsistenz) kann mit dem ‚domain sampling model' durchgeführt werden, nach dem alle aufgestellten Items untereinander korreliert werden und das durchschnittliche ‚r' die Korrelation zu dem gemeinsamen impliziten Summenwert (X_t) bezeichnet (niedrige Inter-Item-Korrelationen werden raus). Eher empfohlen ist Cronbach's alpha, wobei alle Split-Half Variationen eines Bogens durchgetestet werden. Der Endwert alpha (=.5/.6 explorativ; =.9/.95 Standardisierung) gibt das Reliabilitätsmaß an. Fragen werden aussortiert über die Berechnung des alpha's ohne die jeweilige Frage. Eine Faktoranalyse dient zur Ermittlung der Dimensionen von Konstrukten, ist aber erst nach der Aussortierung von Items empfohlen worden, da schlechte Items sonst unnötige Dimensionen erzeugen. Es ist Iterativ vorzugehen, so dass bei Fehlern immer ein Schritt zurück folgt – im Zweifel bis zur Konstrukt-Definition.

6) Die Einflüsse transienter Persönlichkeitsfaktoren sollten mit between-test (also verschiedene Tests die ähnlichen Ergebnisse kommen und nicht mit test-retest Verfahren (gleiche Test nach einer gewissen Pause noch mal durchführen) ermittelt werden. Ursache dafür sieht der Autor in dem Gedächtnis, wodurch diese Ergebnisse per sé beeinflusst werden.

7) Ordentliche Konstrukte und konsequente Reliabilitätsanalysen (für Konsistenz) liefern bereits face und content valide Bögen. Jedoch keine Konstruktvalidität. Dazu wird eine Multi-Trait-Multi-Method Matrix benötigt in der die Hauptdiagonale aus alpha Werten (Cronbach, Konsistenz) und die Unterdiagonalen aus konvergenten Validitäten. Letztere sollen größer sein als alle Korrelationen der multi-trait-mono-method Matrizen und die jeweiligen Korrelationen der Zeilen & Spalten der multi-trait-hetero-method Matrizen. Wenn die Korrelationen der multi-trait-mono-method allgemein niedrig sind, dann kann auf eine gute diskriminante Validität geschlossen werden (VORSICHT ! – TEXT UNKLAR).

8) Wenn mit großen Zahlen gearbeitet wird sollten für Vergleiche einzelner mit den Gesamtscores Mittelwert und Streuung sowie Perzentile berechnet werden. Für die Vergleiche einzelner ist das nicht notwendig.

9) Marketing als Wissenschaft hängt von der Qualität der Instrumente ab. Wer nach dem Vorbild von Churchill arbeitet mach ‚quality research'.

1) Introduction

- Zitiert wird aus Jacoby (1978, Journal of Marketing Research) nachdem die wichtigsten Kriterien für Testverfahren „validity, reliability, and sensitivity“ (Jacoby 1978, S.91 hier nach S. 64) sind. Zudem wird aus den ‚Marketing News' Herr Gardener zitiert, nach dem Sozialwissenschaftler oft Zahlen zählen, aber diese nicht hinterfragen. Summa sumarum: „garbage in, garbage out“ (S. 64).
- Dem gegenüber gilt aber „the obvious need for better measures“ (S. 64).
- Es wird auf die Wissenschaftsgeschichte der Psychologie verwiesen die sich ebenfalls in einer solchen Krise befand (zitiert wird aus Tryon 1957, S. 229). Besonders erfolgreich zur Überwindung dienten dabei nicht die „simple questions of reliability“ (S. 65), sondern direkte und messbare Fragen nach „validity“ (S. 65).
- Ziel dieses Artikels ist es daher ein Framework zur Evaluierung von Marketing-Instrumenten zu entwickeln.

2) The Problem and Approach

- Kern der Operationalisierung ist es „assigning numbers to objects to represent quantities of attributes“ (Nunnally 1967, S. 2 hier nach S. 65). Hierin sieht der Autor zwei wesentliche Punkte: 1) “attributes of objects are measured and not the objects themselves“ (S. 65) und 2) fehlt jedoch dabei jede Regel nach der die Zuschreibung der Zahlen erfolgen soll.
- Problem bei einer jeden Messung (X_0) ist aber, dass zwar der ‚wahre' Wert des Konstruktes erhoben werden soll, der Messwert aber davon abweichen kann. Auf Grund verschiedener Einflussfaktoren (stable characteristics of the person = Charaktereigenschaften, transient personal factors = Persönlichkeitsfaktoren, situational factors = situationsbezogene Merkmale, variation of administration = Erhebungsunterschiede, Sampling of Items = Fragenauswahl, lack of clarity = Unklarheiten in den Fragen, mechanical factors = formenbedingte Probleme wie radieren etc.; S. 65) wir davon ausgegangen, neben dem ‚wahren' Wert (X_t) einen systematischen (X_s) und einen Zufallsfehler (X_r) zu erheben.
- Als Valide gilt, wenn ein Fragebogen „reflect true differences on the characteristic on one is attempting to measure” (S. 65) - also X_0 = X_t. Als Reliabel gilt, wenn “independent but comparable measures of the same trait or construct … agree” (S. 65) – also X_r = 0. Wobei Reliabilität eine Notwendige, aber nicht hinreichende Bedingung für Validität ist ~ “if a measure is valid, it is reliable” (S. 65).
- Problem bei der Analyse von Validität ist (Reliabilität kann durch das alpha bestimmt werden), dass keiner „knows for sure what the X_t scores are“ (S. 66). Somit müssen alle Möglichkeiten der Evaluierung ausgenutzt werden, für Reliabilität (split-half, test-retest; S. 66) wie auch für Validität (face, convergent, discriminant; S. 66).
- Es wird ein Verfahren (Figure 1; S. 66) für „multi-item measures“ (S. 66) vorgeschlagen (Domain of Construct [3.], Items [4.], Data, Analysis [5.]; Data, Reliability [6.], Validity [7.], Standardization).
- Warum Mutli-item Skalen besser sind, liegt daran, dass single item-Skalen aus drei Gründen schlecht sind: 1) „specificity … low corrolation with the attribute“ (S. 66), 2) „categorize people into a relatively small number of groups“ (S. 66), und 3) „same scale position unlikely to be checked“ (S. 66). Lösung ist daher die Verwendung von Multi-Item Skalen die diese Probleme beheben.

3) Specify Domain of the construct

- Jeder Forscher sollte zu Beginn “specifiying the domain of the constrct“ (S. 67) betreiben, was soviel heißt wie „exacting in delineating what is included in the definition and what is excluded“ (S. 67).
- Natürlich sollte bei diesen Definitionen immer schön eine Literatur-Recherche vorausgehen bei der der Forscher ein Statement darüber platziert warum seine Variante nun benötigt wird (S. 67).

4) Generate Sample of Items

- Die Fragen (Items, Item-Batterien), welche zu den jeweiligen Konstrukten gehören sollten durch eine Literaturrecherche (product brochures, articles in trade magazines, and newspapers, or results of product tests), Erfahrungsumfragen (Experteninterviews mit solchen der Produktgruppe, Verkaufsleiter, Verkäufer, Konsumenten, Marketing Researcher, und Outsiders) und „insight-stimulating examples“ (performance letters, a comparison of competitors' products) gewonnen werden (Selltiz et al. 1976 hier nach S. 67). Diese Quellen verraten viel über die Dimensionen / Komponenten von Variablen.
- Weitere Quellen sind “critical incidents and focus groups” (Scenarioerstellung und Bearbeitung mit Schlüsselkonsumenten; S. 67). Im Rahmen dieser Untersuchungen können die Fragen bereits manipuliert bzw. verbessert werden (S. 68). Letzteres besonders bei zweideutigen Fragen, solche mit hohem Grad an sozialer Erwünschtheit oder einfache ja/nein-Fragen.

5) Purify the Measure

INTRODUCTION

- Ältestes Model zur bestimmt des Messwertes (X_t) ist das “domain sampling model” (S. 68) nach dem alle Items verwendet werden um die Gesamtscore der Variable zu ermitteln (Nunnally 1967, S. 175 ff. hier S. 68). Hierbei werden alle Fragen eines Bogens mit allen Fragen korreliert wobei das durchschnittliche r aller Korrelationen untereinander die Korrelation zu einem impliziten ‚commen core' (S. 68; auch Ley 1972) angibt.
- Da aber nicht alle Items aus großen Fragebatterien verwendet werden ist die Frage welche am besten geeignet wären.

COEFFICIENT ALPHA

- Cronbach's alpha (Korrelation aller möglichen Split-half Varianten) misst die “internal consistency of a set of items” (S. 68). Dies wird zunächst als Globalwert berechnet (.50 to .60 für explorative Studien und .90-.95 als Standard nach Nunnally 1967) und dessen Quadratwuzel ist ein Wert für die Korrelation “of the k-item test with errorless true scores“ (S. 68).
- Wenn niedrige alpha Werte gegeben sind, dann sollten Items aussortiert werden (z.B. Korrelation der Items mit dem Gesamtscore; oder alpha ohne jeweiliges Item).
- Bei mehreren Dimensionen einer Variable sollte alpha für jede Dimension separat berechnet werden und die Aussortierung basiert auf der Korrelation zu dem Dimensionsscore (S. 69). Die Gesamt-Reliabilität ist anschließend durch „reliability of linear combinations“ (Nunnally 1967, S. 226 ff. hier nach S. 69) berechnet werden.
- Eine Verwendung eines einfachen Split-half (Auftrennung aller Fragen in gerade und ungerade; Ermittlung der Total-Scores für jeden Block; Korrelation der einzelnen Block-Fragen mit dem Block-Score) ist nicht zu empfehlen, da nur auf 1 Art geprüft wird.

FACTOR ANALYSIS

- Methode zur Ermittlung der “number of dimensions underlying the construct” (S. 69). Es wird jedoch empfohlen die Faktoranalyse nach der Item-Selektion durchzuführen, damit nicht die Garbage-Items unsinnige Dimensionen erzeugen.

ITERATION

- Wenn alpha hoch ist und die Faktoranalyse alle Dimensionen aus der Theorie bestätigt, dann können neue Daten gesammelt werden.
- Wenn im Ergebnis der Faktoranalyse eine Überlappung der Dimensionen angezeigt wird, so sind die jeweiligen Items einzubehalten (S. 69)
- Wenn die Ergebnisse der Faktoranalyse nicht sauber sind und das alpha zu niedrig, dann sollte mit dem Prozess erneut begonnen werden.

6) Assess Reliability with New Data

- Validität (insbesondere face und content) wird durch Experten überprüft und dann gelten Fragen oft als "look right" (S. 69). Auf Grund der erwähnten transienten Persönlichkeitsfaktoren oder Problemen bei Fragen ist es wichtig auch between-test reliability zu analysieren (S. 70).
- Test-retest reliability zu messen wird hier nicht empfohlen wegen dem Erinnerungsvermögen der Probanden (S. 70).

7) Assess Construct Validity

INTRODUCTION

- Durch die exakte Definition und die Generierung von Fragen innerhalb dieser, sowie eine Konsequente Aussortierung von Items durch Reliabilitätsanalysen, kann bereit davon ausgegangen werden, dass die Items „content or face valid" sind (S. 70).
- Aber: „construct validity, which lies at the very heart" (S. 70) kann nun gegeben sein, muss aber nicht "Consistency is necessary but not sufficient for construct validity" (S. 70).
- Zur Berechnung der Validität muss 1) gemessen werden wie hoch die Korrelation zu anderen Instrumenten ist und 2) ob die Ergebnisse die Erwartungen bestätigen. Verwendet wird hierfür die Multi-Trait-Multi-Method-Matrix (MTMM).

CORRELATIONS WITH OTHER MEASURES

- Ausgangspunkt für Vergleichsanalysen von Instrumenten ist, dass das „fundamental principle in science ist hat any particular construct or trait should be measurable by at least two, and preferable more, different methods" (S. 70).
- Die konvergente Konstrukt-Validität (Deckung mit anderen Instrumenten) des Tests ist durch die Höhe der Korrelation mit anderen Methoden (Likert Skala vs. Thermometer vs. Beobachtung) gegeben. Die diskriminante Konstrukt-Validität (nur repliziert oder tatsächlich neu) kann durch die niedrige Korrelation zu sehr ähnlichen Test ermittelt werden (Likert-Skala Hans-Wurst zu Likert-Skala Samuel-Eule).
- Die Korrelationsergebnisse werden in einer MMTM dargestellt (Table 1, S. 71).
    - Die Haupt-Diagonale ist mit den alpha's gefüllt (zwischen .5 und .95).
    - Die Unterdiagonalen sind konvergente Validitäten (Ergebnisse der Methoden-Unterschiede sollen hoch korrelieren).
    - Konvergente Validitäten müssen höher sein als die jeweiligen Korrelationen der Zeile & Spalte in den hetero-method Matrizen.
    - Konvergente Validitäten müssen höher Sein als alle Korrelationen in den mono-method Matrizen.
    - Die Pattern (Vorzeichen und Reihenfolge der Korrelationshöhe je Position) der zusammengehörigen Dreiecke sollten gleich sein.

DOES THE MEASURE BEHAVE AS EXPECTED

- Aus diesem Abschnitt verbleibt nur die Vermutung zu entnehmen, dass niedrige Korrelationen der multi-trait-mono-method Matrizen für eine gute diskriminante Validität sprechen: „analyst tries to establish the construct validity of a measure by relating it to a number of other constructs" (S. 72).
-

8) Developing Norms

- Da durch die Kodierung nicht direkt etwas über den Inhalt oder Wert eines Ergebnisses gesagt werden kann, sollten ‚Normen' oder Vergleichswerte entwickelt werden. Dies wird für Items oder Item-Batterien bzgl. eines Attributes über Mittelwert und Streuung erreicht, mit dem dann individuelle Ergebnisse verglichen werden können (S. 72).
- Wenn nur zwei Personen verglichen werden sollten kann auf den Normungs-Prozess der Ergebnisse verzichtet werden.

9) Summary and Conclusions

- Blabla und natürlich „Progress in the development of marketing as a science certainly will depend on the measures marketers develop to estimate the variables of interest" (Bartels 1951; Buzzell 1963; Converse 1945; Hunt 1976 hier nach S. 73).
- Dieser Prozess ware "quality research" (S. 73)