

Statistik für Informationsmanager

Klausur vom 27. Juli 2005

Name: Max Mustermann

Matrikelnummer:

Erlaubte Hilfsmittel: Außer Schreibwerkzeug diesmal auch Taschenrechner; Antworten bitte im Anschluss an die Fragen, Rückseiten können für Antworten und Entwürfe genutzt werden; Entwürfe bitte anschließend durchstreichen!

Diese Klausur soll — nicht — als Freiversuch gewertet werden.

Mit der elektronischen Veröffentlichung meiner Klausurnote zusammen mit meiner Matrikelnummer bin ich — nicht — einverstanden:

.....
Unterschrift

Für die Teilaufgaben wird es folgende Punktzahlen geben:

Aufgabe	Maximal erreichbar	Im Mittel wurden erreicht	Standardabweichung
1	39	13.8	9.8
2	22	17.4	2.8
3	18	15.5	2.3
4.1	4	3.3	1.6
4.2	10	5.9	1.6
4.3	3	0.7	1.2
4.4	3	0.7	1.0
5.1	2	1.2	0.9
5.2	4	3.8	1.0
5.3	2	1.6	0.7
5.4	8	3.4	3.3
5.5	4	2.3	1.5
5.6	5	2.8	1.5
6.1	4	4.0	0.0
6.2	4	3.4	1.4
6.3	8	5.4	3.2
6.4	2	1.9	0.2
6.5	2	1.8	0.6
Summe	144	83.4	27.5

72 Punkte entsprechen genau einer 4.0, die zweitbeste Arbeit mit 110 Punkten erhielt eine (exakte) 1.3, die übrigen Noten ergaben sich durch entsprechende Berechnung und Rundung auf die nächst benachbarte zulässige Note.

1. Grundlagen

Sie haben in der Vorlesung mehrere multivariate Verfahren kennen gelernt. Jedes von ihnen erfüllt im Forschungsprozess eine oder auch mehrere spezifische Funktionen, die in der unten stehenden Übersicht aufgeführt sind. Markieren Sie in der Übersicht, welchem Verfahren welche Funktionen zukommen und geben Sie zu jeder Markierung einen kurzen Kommentar.

Funktion	Verfahren			
	Multiple Regression	Faktorenanalyse	Clusteranalyse	Diskriminanzanalyse
Hypothesenbildung		1	2	
Hypothesenüberprüfung	3			4
Skalenkonstruktion		5		
Konstruktvalidierung		6		
Vorhersage	7			8
Gruppenbildung			9	
Zuordnung zu Gruppen				10
Segmentierung			11	
Klassifikation				12
Datenreduktion		13		

Meine Kreuzchen begründe ich wie folgt (für jedes gut begründete Kreuzchen gibt es 3 Punkte, für jedes richtige, aber nicht begründete Kreuzchen gibt es einen Punkt, für jedes falsche Kreuzchen ohne oder mit nicht überzeugender Begründung wird ein Punkt abgezogen — maximal kann es 39 Punkte geben! Am besten nummerieren Sie die Kreuzchen durch und benutzen die Nummern für Ihre Begründungen):

1 Hypothesen über den Zusammenhang zwischen Variablen, über latente Variable,

2 Hypothesen über Segmente, Gruppen etc.

3 Überprüfung von Hypothesen über Zusammenhänge zwischen Variablen

4 Überprüfung von Hypothesen über Zusammenhänge zwischen einer Gruppierungs- und mehreren metrischen Variablen

5 Konstruktion einer oder mehrerer Skalen für latente Variable

6 Überprüfung, ob Items einer Fragebatterie zu einer einzigen Dimension gehören

7 Vorhersage von künftigen Messwerten der abhängigen Variablen bei einem zusätzlichen Fall, wenn dessen Werte in den unabhängigen Variablen bekannt sind

8 Vorhersage der Gruppenzugehörigkeit bei einem zusätzlichen Fall, wenn dessen Werte in den unabhängigen Variablen bekannt sind

9 Auffinden von Gruppen, Clustern

10 wie 8

11 wie 9

12 wie 10

13 Bilden weniger Skalen aus vielen Indikatoren

Es gab auch für andere Kreuzchen Punkte, wenn die Begründungen einigermaßen überzeugend waren.

2. Skalenniveaus

Geben Sie jeweils für die folgenden Variablen an, ob sie nominal, ordinal, intervall- oder ratioskaliert sind (jeweils ein Punkt) und geben Sie jeweils eine kurze Begründung (jeweils ein weiterer Punkt). (Eine gute Begründung für eine unserer Ansicht nach falsche Einordnung gibt eventuell ebenfalls Punkte, falsch gesetzte Kreuzchen ohne oder mit nicht überzeugender Begründung führen zu einem Punktabzug! Maximal sind 22 Punkte möglich, denn in einigen Fällen lassen sich zwei Kreuzchen pro Zeile gut begründen.)

	Nominal	Ordinal	Intervall	Ratio	Begründung
Aktienindex				X	Division zweier Indexwerte ist sinnvoll
Sympathieskalometer (mit einer Skala von -5 bis +5)		X	X		Alle Abstände können evtl. als gleich angesehen werden, kein nat. Nullpunkt
Inflationsrate				X	„Die Inflationsrate hat sich gegenüber dem Vorjahr verdoppelt“ ist eine sinnvolle Aussage.
Farbe	X	X			Keine eindeutige eindimensionale Anordnung möglich, es sei denn man beschränkt sich auf die „reinen“ Farben
Bezeichnung eines Autotyps	X				dito
Rangplatz bei einem Wettkampf		X			Abstände können nicht als gleich angesehen werden, Ordnung aber möglich
Lottozahl	X				... ist nur der Name einer Zahl, keinerlei Rechenoperationen sinnvoll
Klausurnote		X	X		Evtl. können alle Abstände als gleich angesehen werden.

3. Univariate Statistik

Berechnen Sie (jeweils falls sinnvoll!) den Modus, den Median und den Mittelwert für folgende Variablen (richtige Berechnung[en] jeweils 1P, richtige Begründung jeweils noch 1P, wenn bei einer Teilaufgabe richtigerweise alle drei Werte berechnet worden sind, bedarf es keiner Begründung, und es gibt trotzdem 2P):

1. Inflationsrate aufeinander folgender Jahre (Preissteigerung in %, 1965 gegenüber 1964 bis 1989 gegenüber 1988):

2.1	2.0	6.4	-0.2	1.1	3.1	4.7	4.6	6.6	7.2
5.4	3.2	3.9	2.4	3.5	5.2	5.2	4.8	2.7	2.0
Mittelwert:	3.495; ist aber nicht sehr sinnvoll, was aber allenfalls eine(r) geahnt hat!								
Median	Alles zwischen 3.2 und 3.5								
Modus	Ca. bei 2.1 und bei 5 (2.0 und 5.2, weil sie jeweils zweimal vorkommen, wurde								
gegebenenfalls Begründung für nicht berechnete Parameter:	ebenfalls als richtig bewertet, 2.1 und 5.0 dürften jedoch näher an „wahren“ Werten liegen.								

2. Links-rechts-Einstellung auf einer Skala von 0 („ganz links“) bis 10 („ganz rechts“) — Angaben von 20 repräsentativ ausgewählten luxemburgischen Befragten aus dem Eurobarometer 2003

1	2	2	3	3	4	4	4	5	5
5	5	5	5	5	6	6	7	8	10
Mittelwert:	4.75								
Median	5								
Modus	5								
gegebenenfalls Begründung für nicht berechnete Parameter:									

3. Region, in der die zur Zeit am Campus Studierenden ihre Hochschulzugangsberechtigung erworben haben:

Region	Koblenz	MYK/EMS /WW	Rest des ehem. Reg.-Bez Koblenz	Rest von Rheinland-Pfalz	NRW, Hessen, Baden-Württ., Saarland	Rest von Deutschland	Ausland
Anzahl	834	1089	1007	790	1340	388	74
Mittelwert:							
Median							
Modus	NRW etc.						
gegebenenfalls Begründung für nicht berechnete Parameter:	Weder ordinal noch metrisch						

4. Bivariate Statistik

Im Eurobarometer 2003 wurden die Befragten nach ihrem Beruf befragt. Hier sind die Antwortverteilungen der größten EU-Staaten (einige Berufskategorien wurden zusammengefasst):

	COUNTRIES (NATION II)					Total
	DEUTSCHL AND WEST	ITALIA	ESPANA	FRANCE	GREAT BRITAIN	
1.Farmer, Fishermen	19 .6%	12 .5%	5 .3%	23 .9%	5 .2%	64 .5%
3.Professionals	97 3.3%	135 5.4%	48 2.8%	32 1.3%	53 2.2%	365 3.0%
4.Owner of a shop, craftsmen, self employed	130 4.4%	234 9.3%	155 8.9%	153 6.0%	102 4.3%	774 6.4%
7.General and middle management	283 9.5%	154 6.1%	64 3.7%	269 10.5%	135 5.7%	905 7.5%
9.Employed position	539 18.1%	456 18.2%	316 18.2%	649 25.4%	369 15.6%	2329 19.2%
13.Skilled manual worker	254 8.5%	122 4.9%	209 12.1%	231 9.1%	194 8.2%	1010 8.3%
14.Other (unskilled) manual worker, servant	121 4.1%	100 4.0%	116 6.7%	43 1.7%	259 11.0%	639 5.3%
15.Responsible for looking after the household	287 9.7%	310 12.4%	273 15.7%	262 10.3%	385 16.3%	1517 12.5%
16.Student	242 8.1%	263 10.5%	173 10.0%	185 7.2%	152 6.4%	1015 8.4%
17.Unemployed or temporarily not working	143 4.8%	106 4.2%	55 3.2%	128 5.0%	170 7.2%	602 5.0%
18.Retired or unable to work through illness	858 28.9%	617 24.6%	320 18.5%	577 22.6%	536 22.7%	2908 24.0%
Total	2973 100.0%	2509 100.0%	1734 100.0%	2552 100.0%	2360 100.0%	12128 100.0%

1. Handelt es sich bei den beiden Variablen um nominal oder um ordinal skalierte Variable? (4 Punkte)

Beide sind nominal.

2. Formulieren Sie zu diesem Zusammenhang zwei bis drei Sätze! Machen Sie Unterschiede in der Berufsstruktur in den fünf „großen“ EU-Staaten deutlich. (bis zu 10 Punkte)

Hier kam es mir darauf an, die wesentlichen Unterschiede zwischen den Ländern zu identifizieren; die Punktzahl richtete sich nach der Ausführlichkeit der mitgeteilten Beobachtungen.

3. Für diesen Zusammenhang ist Phi 0.264 und Cramers' V 0.132. Erklären Sie den Unterschied dieser beiden eng verwandten Maße. (3 Punkte)

Cramers' V ist ein quadratischer Wert, außerdem ist er für die Zeilen- und Spaltenzahl korrigiert (deswegen ist er größer als das Quadrat von Phi).

4. Die PRE-Maße (Beruf „abhängig“) sind Lambda = 0.008, Goodman's and Kruskal's Tau = 0.017, Uncertainty Coefficient = 0.0016. Erklären Sie, warum diese Maße kleiner sind als die von Chi-Quadrat abgeleiteten. (3 Punkte)

Die PRE-Maße sagen numerisch etwas aus über den Ratefehler — der bleibt auch bei Kenntnis der Landeszugehörigkeit groß, wird also nur geringfügig vermindert. Die von Chi-Quadrat abgeleiteten Werte (Phi und Cramers' V) wurden meist gar nicht angesprochen, dafür wurde häufig gesagt, dass Chi-Quadrat von der Größe der Stichprobe ist, was zwar richtig ist, aber keine Antwort auf diese Frage war.

5. Multivariate Statistik / Multiple Regression

Für alle (bisher 174) Studierenden des Bachelor-Studiengangs Informationsmanagement lässt sich — auch wenn sie den Abschluss noch nicht erreicht haben — eine vorläufige Gesamtnote aus den vorliegenden Leistungsdaten berechnen. Sie kann als abhängige Variable in einer Regression auf einige wichtige Klausuren der ersten beiden Semester untersucht werden, etwa um Studienanfängern frühzeitig ihre Erfolgsaussichten zu verdeutlichen. Die folgende Übersicht gibt die wichtigsten Koeffizienten wieder:

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	.900	.095		9.478	.000
BWLNote	.124	.026	.261	4.708	.000
InfNote	.099	.028	.202	3.594	.000
MathNote	.199	.035	.316	5.654	.000
StatNote	.057	.026	.121	2.155	.033
VWLNote	.109	.028	.234	3.910	.000

a Dependent Variable: gesamtnote

5.1 Was bedeuten die beiden Zahlen **0.900** und **0.124** in den ersten beiden Zeilen der Tabelle? (2P).

0.900 wäre die zu erwartende Gesamtnote, wenn alle Einzelnoten genau 0.0 wären; 0.124 ist der Unterschied zwischen den zu erwartenden Gesamtnoten zweier Studierenden, die sich ausschließlich in der BWL-Note um genau 1.0 unterscheiden.

5.2 Berechnen Sie die zu erwartende Gesamtnote eines/r Studierenden, der/die in Informatik I für Informationsmanager (InfNote) eine 2.0 und in den anderen Klausuren jeweils eine 3.0 erreicht hat. (4 P).

$$2.565 = 0.9 + 3 (0.124 + 0.199 + 0.057 + 0.109) + 2 * 0.99$$

5.3 Aus welcher der fünf Noten lässt sich am besten, aus welcher am wenigsten gut auf die spätere Gesamtnote schließen? (2 Punkte) Wie können Sie sich dieses Ergebnis erklären? (noch einmal 8 Punkte)

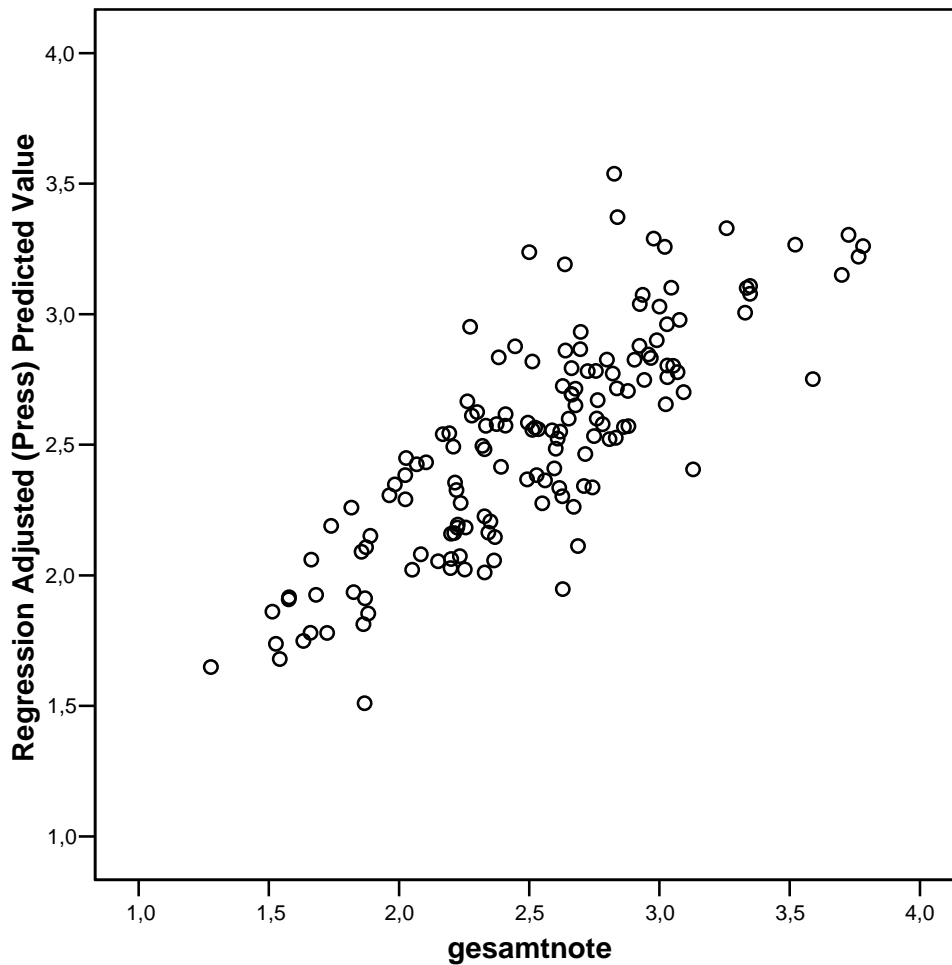
Aus Mathematik am besten, aus Statistik am schlechtesten. Die beste Antwort auf den zweiten Teil der Frage war: „... In der Statistik geht es aber nicht primär um Rechnungen, sondern um die Fähigkeit, Daten zu analysieren, also nicht um Mathematik primär, sondern um Analysefähigkeit ...“

5.4 Drücken Sie in einem ganzen Satz aus, was die Zahl **0.033** (am Ende der vorletzten Spalte) bzw. ihr Kehrwert 30 bedeutet! (4 P)

Die Wahrscheinlichkeit, in einer gleich großen Stichprobe aus der gleichen Grundgesamtheit einen Koeffizienten > 0.057 (bzw. einen standardisierten Koeffizienten > 0.121) zu finden, wenn der entsprechende Koeffizient in der Grundgesamtheit gerade 0.0 ist, beträgt 0.33 oder 3.3 %.

Scatterplot

Dependent Variable: gesamtnote



5.5 Was entnehmen Sie dem Streuungsdiagramm, das in der Waagerechten die tatsächlichen (bisherigen, vorläufigen) Gesamtnoten und in der Senkrechten die Werte enthält, mit der die künftige Gesamtnote aus den fünf Einzelnoten der ersten beiden Semester vorhergesagt wird? (5 Punkte)

Hier war etwas über die Linearität des Zusammenhangs und über die Homoskedasizität zu sagen, eventuell auch darüber, dass die Schätzfehler denn doch bei ± 1 liegt.

6. Multivariate Statistik / Faktorenanalyse

Im Eurobarometer 2003 wurden neben den Fragen zur Internetnutzung auch mehrere Gruppen von Fragen zum Thema Sport gestellt. In der Tabelle unten finden Sie die Faktorladungen der Einzelantworten auf drei Fragebatterien. Q39 wurde eingeleitet mit dem Fragetext „Was sind die Hauptvorzüge von Sport?“, bei Q40 lautete die Einleitung „Welche der folgenden Werte werden durch den Sport am meisten gefördert?“ und bei Q41 war sie „Bitte sagen Sie zu jeder der folgenden Aussagen, ob Sie zustimmen oder nicht!“. Bei Q41 gab es nur die Antwortmöglichkeiten „stimme zu“ bzw. „stimme nicht zu“, bei den beiden anderen steht in der Datei eine 1 für jeden genannten Vorzug oder Wert, sonst eine 0.

Koeffizienten kleiner als 0.11 sind zur Erleichterung des Auffindens einer Einfachstruktur weggelassen, außerdem sind die Variablen nach ihren höchsten Faktorladungen sortiert. Interpretieren Sie die beiden Faktoren, indem Sie ihnen möglichst prägnante Namen geben. Begründen Sie Ihre Namensgebung.

Rotated Component Matrix(a)

	Component	
	1	2
Q40 SPORT VALUES: MUTUAL UNDERSTANDING	.593	
Q40 SPORT VALUES: SOLIDARITY	.579	
Q39 SPORT BENEFIT: BUILD CHARACTER	.579	
Q40 SPORT VALUES: EQUALITY	.579	
Q39 SPORT BENEFIT: DEVELOP NEW SKILLS	.561	
Q39 SPORT BENEFIT: INTEGR DISADVANTAGED	.556	
Q39 SPORT BENEFIT: IMPROVE SELF-ESTEEM	.545	
Q39 SPORT BENEFIT: MEET OTHER CULTURES	.531	
Q39 SPORT BENEFIT: ACHIEVE OBJECTIVES	.531	
Q39 SPORT BENEFIT: NEW ACQUAINTANCES	.527	
Q40 SPORT VALUES: RESPECT FOR OTHERS	.521	
Q40 SPORT VALUES: STICKING TO RULES	.509	
Q40 SPORT VALUES: FAIR PLAY	.501	
Q39 SPORT BENEFIT: SPIRIT OF COMPETITION	.484	
Q39 SPORT BENEFIT: BE WITH FRIENDS	.479	
Q40 SPORT VALUES: TOLERANCE	.468	-.105
Q40 SPORT VALUES: FRIENDSHIP	.468	
Q40 SPORT VALUES: SELF-CONTROL	.444	
Q40 SPORT VALUES: DISCIPLINE	.421	
Q40 SPORT VALUES: EFFORT	.403	
Q39 SPORT BENEFIT: HAVE FUN	.385	
Q39 SPORT BENEFIT: PHYSICAL PERFORMANCE	.363	
Q40 SPORT VALUES: TEAM SPIRIT	.340	
Q41 SPORT: EU should more actively promote education through sport		.717
Q41 SPORT: EU should co-operate more with national sports organisations and national governments		.713
Q41 SPORT: EU should be able to intervene more in European sports issues		.699
Q41 SPORT: The promotion of the ethical and social values of sport should be a priority for the EU		.683
Q41 SPORT: There should be better co-operation between educational institutions and sports organisations in (your country)		.576
Q41 SPORT: Through sport you can fight against any form of discrimination		.561
Q41 SPORT: More time should be devoted to sport in school timetables		.539
Q41 SPORT: Sport promotes the dialogue between cultures		.489
Q41 SPORT: Sports makes it easier to fight sedentary habits		.470
Q41 SPORT: It is easy to find a balance between sport and other leisure activities		.463
Q41 SPORT: EU should participate in the fight against doping		.437

6.1 Faktor 1 beschreibt die latente Eigenschaft (Dimension, Einstellung der Befragten) ... (4 P)

..., auf die Fragen Q39 und Q40 einheitlich zu antworten ...

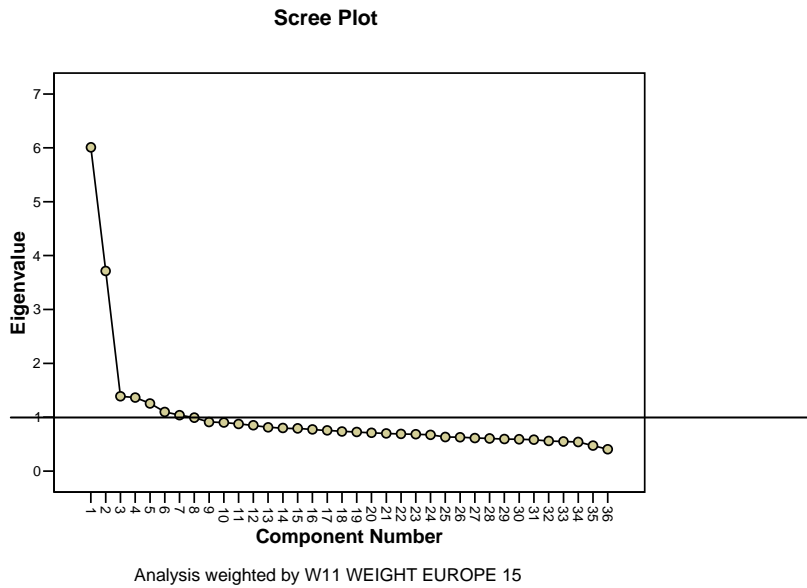
6.2 Faktor 2 ... (4 P)

... repräsentiert dann die Eigenschaft, auf die Frage Q41 einheitlich zu antworten.

6.3 Wie beurteilen Sie die Einfachstruktur? Besteht Grund zu der Annahme, dass es sich hier um ein Artefakt halten könnte (dass die Einfachstruktur das Ergebnis einer ungeschickten Fragebogenkonstruktion sein könnte)? (8 P)

Die Einfachstruktur ist gut bis exzellent, aber das kann einfach an der Art der jeweiligen Fragestellung liegen: Es fällt den Befragten bei diesen Fragebatterien offenbar schwer, zwischen den Einzelstatement zu differenzieren.

Hier finden Sie den Scree-Plot zu der vorstehenden Auswertung:



Halten Sie danach die Extraktion von zwei Faktoren noch für gerechtfertigt? (Ja / Nein / Kommt darauf an)

Wenn ja: warum? (2 P)

Der Knick bei Faktor 3 ist ziemlich unübersehbar, die anderen Knoten liegen praktisch genau auf einer Geraden.

Wenn nein: warum nicht? Wie viele hätte man statt dessen extrahieren sollen? (2 P)

Nach dem Kaiser-Kriterium hätte man 6–8 Faktoren genommen (genauer ist das in der Abb. nicht zu erkennen); aus dem Scree-Test wäre auch noch eine Faktorenzahl von 5 zu erkennen, denn die Knoten ab 6 liegen noch besser auf einer Geraden.

(Antworten auf beide Unterfragen werden gewertet! Wenn Sie dezidiert der Auffassung sind, dass nicht zwei, sondern einer oder drei oder vier oder fünf oder sechs oder sieben oder acht ... Faktoren hätten extrahiert werden sollen, müssen Sie auf „wenn nein: warum nicht?“ besonders ausführlich antworten; ebenso müssen Sie, wenn Sie dezidiert der Auffassung sind, dass drei Faktoren die richtige Wahl waren, auf „wenn ja, warum?“ besonders ausführlich antworten.)