„Don't make me think aloud!" – Lautes Denken mit Eye Tracking auf dem Prüfstand

Was wir von der Methode des retrospektiven Lauten Denkens lernen können

Yong-Min Markus Jo
Immobilien Scout GmbH
Andreasstraße 10, 10243 Berlin
markus.jo@immobilienscout24.de

Anke Stautmeister
Immobilien Scout GmbH
Andreasstraße 10, 10243 Berlin
anke.stautmeister@immobilienscout24.de

Abstract
Das Laute Denken während der Aufgabendurchführung in Usability-Tests (Concurrent Think Aloud, kurz: CTA) ist reaktiv. Die Erfassung des natürlichen Nutzerverhaltens ist nicht möglich. Nachgelagertes Lautes Denken (Retrospective Think Aloud, kurz: RTA) hingegen sollte den gleichen Erkenntnisgewinn bei geringerer Beeinflussung hervorrufen. 40 Probanden absolvierten in vorliegender Studie dieselbe Such-Aufgabe auf einem Immobilienportal, die eine Hälfte mit CTA, die andere mit RTA. Die Ergebnisse wurden mittels acht Hypothesen verglichen. Beim CTA verlängert sich die Aufgabendurchführung signifikant, die Anzahl der Fixationen erhöht sich. Leichte Trends in erwarteter Richtung ergeben sich für die Anzahl betrachteter Bereiche und die Interaktion mit dem Stimulus. Mit CTA werden deutlich mehr pro-aktive Verbesserungsvorschläge geäußert und mehr Usability-Probleme im Bereich „Layout" identifiziert. RTA gewährleistet eine natürlichere Nutzungssituation, CTA bietet eine tiefergehende Untersuchung (z. B. Generierung neuer Features und Konzepte). Je freier die Usability-Aufgabe gestellt wird, desto stärker ist ein Anstieg an Reaktivität des CTA zu erwarten.

Keywords:
/// Usability-Test
/// Methoden-Evaluation
/// Retrospektives Lautes Denken
/// Eye Tracking
/// Gaze Replay

1. Einleitung

Das Laute Denken während einer Aufgabendurchführung (Concurrent Think Aloud, kurz: CTA) in Usability-Tests ist eine Standard-Erhebungsmethode, die nur selten hinterfragt wird. Dabei wurde bereits 1977 vor deren Reaktivität gewarnt: Die Anwendung von CTA kann das Verhalten des Probanden in der Testsituation beeinflussen und so die Ergebnisse verzerren (Nisbett & Wilson 1977). Die Praxis zeigt zudem, dass einige Probanden durch den Moderator fortwährend zum Lauten Denken animiert werden müssen, da dies eine ungewöhnliche Aufgabe darstellt. Jeder Animationsversuch wiederum erhöht die Gefahr der Reaktivität des Lauten Denkens (Ericsson & Simon 1984).

Erschwerend kommt hinzu, dass aus ökonomischen Gründen Usability-Tests nicht selten parallel sowohl mit CTA als auch mit Blickbewegungsmessung (Eye Tracking) durchgeführt werden. Hier ist die Reaktivität offensichtlich, nämlich die Beeinflussung des Blickverhaltens durch das Laute Denken selbst: Wer kommentiert, schaut dabei wahrscheinlich länger auf Bereiche des Stimulus, denn die Aussprache eines Gedankens dauert länger als der Gedanke selbst.

Ein möglicher Lösungsweg für dieses methodologische Problem des CTA, ob mit oder ohne Eye Tracking, ist das nachgelagerte Laute Denken (Retrospective Think Aloud, kurz; RTA): Den Probanden wird direkt im Anschluss an die Aufgabendurchführung ein Video gezeigt, welches die Verhaltensweisen der Person sowie den Stimulus wiedergibt. Bei Web-Usability-Studien kann das z. B. eine Bildschirmaufzeichnung mit der Anzeige der Mausbewegungen und -klicks sein. Der Proband wird also im Nachgang gebeten, laut zu kommentieren, was ihm bei der Aufgabendurchführung durch den Kopf gegangen ist.

Weitere Bedenken werden in der Literatur zur Validität der Methode des Lauten Denkens geäußert, die durch Fabrikation zusätzlicher (Gedanken-)Inhalte oder durch Auslassung existierender Inhalte reduziert wird (Ericsson & Simon 1984). Um diesen Fallstricken zu begegnen hat sich der sogenannte Gaze Replay beim RTA als hilfreich erwiesen: Probanden wird zusätzlich zu den Mausbewegungen und -klicks der eigene Blickverlauf angezeigt, der als zusätzliche Gedächtnisstütze (sog. cue) dienen soll. Der Blickverlauf wird dafür während der Aufgabendurchführung mittels Eye Tracking erfasst und danach ebenfalls im Video abgespielt. Dieses Versuchsdesign, RTA mit Gaze Replay, erschien den Autoren bisher eine valide Methode zur Untersuchung objektiver und subjektiver Variablen eines Web-Usability-Tests zu sein.

Da der Aufwand für die Vorbereitung einer Studie mit Eye Tracking (z. B. umfangreiches Hardware- und Software-Setup) und RTA (z. B. längere Versuchsdauer pro

Proband im Vergleich zu einem Usability-Test mit CTA) wesentlich höher ist, stellte sich jedoch die Frage, ob in der Berufspraxis die Vorteile des RTA mit Gaze Replay so deutlich sind, dass der zusätzliche Aufwand in Kauf genommen und der klassische Usability-Test mit CTA abgelöst werden sollte. Ziel dieser Studie ist es daher, den möglichen Mehrwert von RTA durch einen systematischen Vergleich von CTA und RTA über verschiedene Faktoren zu untersuchen.

2. Hypothesen

Die Autoren erwarteten Auswirkungen auf die Blickbewegungen der Probanden wie auch Auswirkungen auf das Verhalten der Person in der Testsituation. Im Weiteren wird näher auf die einzelnen Hypothesen und deren Grundlagen eingegangen.

2.1. Blickbewegung (Eye Tracking)

Die Blickbewegungsmessung erlaubt die Erhebung von objektiven Verhaltensdaten. Sakkaden („Blicksprünge") und Fixationen („Blickpunkte") der Personen werden nach ihrer Lage und Länge erfasst und können anschließend ausgewertet werden. Einige der Auswertungs-Parameter fließen in die folgenden Hypothesen mit ein.

2.1.1. Hypothese 1: „Sample Rate"

Der Erfassungsgrad der Blicke eines Probanden in % (Sample Rate) fällt bei Anwendung des CTA niedriger aus als bei Anwendung des RTA.

Begründung:
Der Proband neigt beim Lauten Denken während einer Aufgabendurchführung (CTA) dazu, mit dem Moderator zu kommunizieren und ihn dabei anzuschauen. Durch das Anschauen des Moderators gelangen die Pupillen der Probanden aus dem Erfassungsraum des Blickbewegungsmessgerätes. Dadurch werden weniger Blickbewegungsdaten aufgezeichnet; die Sample Rate sinkt.

2.1.2. Hypothese 2: „Total Fixation Duration"

Die Gesamtfixationsdauer (Total Fixation Duration) ist unter Anwendung des CTA länger als unter Anwendung des RTA.

Begründung:
Da Probanden während des CTA laut aussprechen sollen, was ihnen durch den Kopf geht und das Aussprechen eines Gedankens länger dauert als der Gedanke selbst, wird eine längere Gesamtfixationsdauer erwartet; allgemein gesprochen schauen sich Probanden den Stimulus also länger an, während sie erläutern.

2.1.3. Hypothese 3: „Fixationen"

Die Anzahl der erfassten Fixationen („Blickpunkte") innerhalb des Stimulus ist beim CTA höher als beim RTA.

Begründung:
Basierend auf der vorherigen Hypothese sollte sich neben der Gesamtfixationsdauer auch die Anzahl der Einzelfixationen aufgrund einer längeren Betrachtung des Stimulus während des CTA erhöhen. Dies hängt mit einer längeren Gesamtbetrachtungsdauer zusammen, in der sich mehr Möglichkeiten für Fixationen ergeben.

2.1.4. Hypothese 4: „Areas of Interest"

Die Anzahl unterschiedlicher betrachteter Bereiche (Areas of Interest, kurz: AOI) innerhalb des Stimulus (z. B. einer Webseite) ist unter Anwendung des CTA höher als unter Anwendung des RTA.

Begründung:
Der Stimulus kann je nach Interesse der Untersuchung in verschiedene „Areas of Interest" (AOI) unterteilt werden. Diese dienen der Blickerfassungssoftware daraufhin als Auswertungsgrundlage für verschiedene Parameter wie die Anzahl der betrachteten AOI durch einen Probanden. Über die Hypothesen „Total Fixation Duration" und „Fixations" hinaus ist zu erwarten, dass während der bereits diskutierten reaktiv verlängerten Betrachtungsdauer neben den weiteren Fixationen auch mehr Bereiche des Stimulus exploriert werden. Bei der verbalen Erläuterung der Gedanken sollte der Blick also weiter „umherschweifen".

2.2. Verhalten in der Testsituation

Auswirkungen auf das Verhalten einer Person durch CTA, welches sich wiederum auf die Testsituation selbst auswirkt, wird in der Literatur umfassend diskutiert (z. B. Russo, Johnson & Stephens 1989). Im Folgenden werden weitere Hypothesen vorgestellt, welche die Reaktivität des CTA betreffen.

2.2.1. Hypothese 5: „Verweildauer"

Die Verweildauer auf einem Stimulus ist unter Anwendung des CTA länger als unter Anwendung des RTA.

Begründung:
Da Probanden während des CTA laut aussprechen, was ihnen durch den Kopf geht, und der ausgesprochene Gedanke mehr Zeit benötigt als der Gedanke selbst, wird unter dieser Bedingung eine längere Verweildauer auf dem Stimulus erwartet.

2.2.2. Hypothese 6: „Verhalten"

Die zu beobachtenden Interaktionen der Probanden mit dem Stimulus unterscheiden sich abhängig von der Anwendung des CTA oder des RTA.

Begründung:
CTA beeinflusst das Verhalten der Probanden, da eine zusätzliche, kognitiv aufwendige Aufgabe neben der primären Usability-Test-Aufgabe gestellt wird, nämlich die Verbalisierung der Gedanken (Russo, Johnson & Stephens 1989). Die Hypothese macht keine Aussage über die Richtung der Unterschiedlichkeit.

2.2.3.
Hypothese 7: „Usability-Probleme“

Bei der Anwendung von CTA werden mehr Usability-Probleme identifiziert als bei der Anwendung von RTA.

Begründung:
Da sich die Probanden aus der Gruppe „CTA“ bereits während der Interaktion intensiver mit dem Stimulus auseinandersetzen (z. B. längere Verweildauer, vgl. oben), haben sie somit auch länger die Möglichkeit, auf Usability-Probleme zu stoßen. Zudem argumentierten Russo, Johnson und Stephens (1989), dass das Multi-Tasking von Lautem Denken und Aufgabendurchführung zu einer höheren kognitiven Belastung führt, welche häufigere Bedienungsfehler (Usability-Probleme) zur Folge hat.

2.2.4.
Hypothese 8 „Verbale Äußerungen“

Es werden mehr spontane Äußerungen getätigt, wenn CTA angewendet wird.

Begründung:
Durch das Laute Denken während einer Aufgabendurchführung (CTA) entstehen ggf. weiterführende Gedanken wie z. B. spontane Verbesserungsvorschläge, die ebenfalls ausgesprochen werden. Beim RTA dagegen wird sich der Proband eher an die Prozessdauer der Aufgabendurchführung halten, d.h. beim nachträglichen Vorspielen des Videos wird sich der Proband in den Verbalisierungen seiner Gedanken strikter an den ihm gezeigten Verhaltensweisen orientieren. Dadurch gehen beim RTA spontane Äußerungen verloren, die ggf. geholfen hätten, die Usability-Probleme genauer zu umreißen.

3.
Methode
3.1.
Studiendesign

Um die Auswirkungen von CTA und RTA vergleichen zu können wurden zwei Versuchsgruppen erstellt:

Gruppe A (CTA): Lautes Denken während der Aufgabendurchführung

Gruppe B (RTA): Nachgelagertes Lautes Denken während des Vorspielens des Gaze Replays (Video mit Eye Tracking)

Alle Probanden waren Mietwohnung-Suchende aus Berlin, die das zu testende Immobilien-Portal www.immobilienscout24.de kannten. Gruppe A (CTA) bestand aus 20 Personen mit einem Altersdurchschnitt von 31 Jahren, davon waren 9 Teilnehmer weiblich und 11 männlich. Gruppe B (RTA) bestand ebenfalls aus 20 Personen, der Altersdurchschnitt betrug hier 33 Jahre, 13 Teilnehmer waren weiblich, 7 männlich.

Gruppe B (RTA) kann als Kontrollgruppe (Silent Condition) für die Bedingung CTA angesehen werden, da sie das „natürlichere Verhalten“ ohne die Beeinflussung durch die Aufforderung zum Lauten Denken darstellt.

3.2.
Untersuchungsbedingungen

Die Studie wurde innerhalb eines Zeitraums von 3 Wochen im November und Dezember 2010 durchgeführt. In einem separaten Raum des Bürogebäudes der Firma Immobilien Scout GmbH in Berlin wurde ein Dell-Notebook (Latitude E6400) und der Eye Tracker T120 der Firma Tobii Technology installiert. Mit der zugehörigen Aufnahme- und Auswertungssoftware Tobii Studio 2.1 konnten die Probanden und ihr Verhalten in Form von Mausklicks, Mausbewegungen, Blicken und Kommentaren aufgezeichnet werden. Testobjekt war die Webseite www.immobilienscout24.de. „Areas of Interest“ wurden nach gängigen Webseiten-Elementen wie z. B. Logo, Navigation, Menü, Suchfeld usw. definiert.

3.3.
Durchführung

Alle Probanden aus beiden Versuchsgruppen wurden gebeten, nach einer Wohnung zur Miete zu suchen, und zwar in einem von ihnen selbst ausgewählten Stadtteil mit Angabe der monatlichen Netto-Kaltmiete, der Wohnfläche sowie der Anzahl der Zimmer. Neu für alle Probanden war der Sucheinstieg auf der Startseite von ImmobilienScout24, der erstmalig eine Suchmaske darbot, in die alle oben genannten Kriterien sofort eingegeben werden konnten, woraufhin nach Klick auf den Suchen-Button die Treffer direkt erscheinen (bis dato verlief die Eingabe der Suchkriterien über mehrere Seiten). Im Vorfeld wurden alle Probanden über die Video-Aufnahme sowie über die Aufzeichnung der Blickbewegung aufgeklärt.

Folgende Instruktion erhielten alle Probanden aus der Gruppe A (CTA) vor Beginn der Suche: „Bitte äußern Sie alles, was Ihnen durch den Kopf geht, selbst wenn es Ihnen abwegig erscheint. Uns hilft es, wenn Sie viel erzählen.“ Da bekannt ist, dass nicht alle Probanden durchgehend das Laute Denken berücksichtigen, wurde nach jeweils 5 Sekunden des Schweigens jeder Proband mit folgender Frage zum Lauten Denken animiert: „Was geht Ihnen gerade durch den Kopf?“

Hingegen wurde keiner der Probanden aus Gruppe B (RTA) gebeten, während der Aufgabendurchführung laut zu denken. Direkt im Anschluss wurde ihnen das Video (Screen Recording ohne Ton- und Portrait-Aufnahme des Probanden, dafür mit Mausbewegungen und -klicks sowie Blickbewegungen) gezeigt und die Probanden wurden folgendermaßen instruiert: „Bitte äußern Sie alles, was Ihnen vorhin durch den Kopf gegangen ist, selbst wenn es Ihnen abwegig erscheint. Uns hilft es, wenn Sie viel erzählen.“

Ein Pretest und die Erfahrung mit vorausgegangenen RTA-Studien hatte dabei gezeigt, dass es sinnvoll ist, vor Abspielen des Gaze Replays eine kurze Erläuterung über den sichtbaren Blickpunkt (= Fixation) zu liefern, der größer wird, je länger man eine Stelle betrachtet, und dass die schnelle, sprunghafte Bewegung der Augen ein normales Blickverhalten darstellt. Damit die Probanden genügend Zeit für das retrospektive Laute Denken hatten, wurde die Abspielgeschwindigkeit des Gaze Replays um die Hälfte reduziert,

	Mittelwert Gruppe A (CTA) N=20	Mittelwert Gruppe B (RTA) N=20
H1 „Sample Rate“	77,85 % (Range von 21 bis 92 %)	79,85 % (Range von 42 bis 96 %)
H2 „Total Fixation Duration“ (innerhalb der ersten 45 Sek.)	26,12 Sek.* (Range von 5,23 bis 39,86 Sek.)	20,02 Sek.* (Range von 2,89 bis 38,64 Sek.)
H3 „Fixations“ (innerhalb der ersten 45 Sek.)	62,65 * (Range von 9 bis 117)	46,45 * (Range von 2 bis 89)
H4 „Areas of Interest“ (max. Anzahl: 10)	2,85 AOI (Range von 1 bis 5)	2,35 AOI (Range von 1 bis 4)

Tab. 1.
Ergebnisse der Eye-Tracking-Daten getrennt nach Versuchsbedingungen
*signifikant auf dem 5%-Niveau

	Gruppe A (CTA) N=20	Gruppe B (RTA) N=20
H5 „Verweildauer“ (auf der Startseite, Mittelwert)	42,63 Sek.* (Range von 5,23 bis 82,57 Sek.)	27,08 Sek.* (Range von 3,73 bis 39,1 Sek.)
H6 „Verhalten“ (alternativer Sucheinstieg gewählt)	3	6
H7 „Usability-Probleme“ (inkl. Dopplungen)	41	48
H7 „Usability-Probleme“ (ohne Dopplungen)	24	20
H8 „Äußerungen“ (Anzahl pro-aktiver Verbesserungsvorschläge, inkl. Dopplungen)	13	1
H8 „Äußerungen“ (Anzahl pro-aktiver Verbesserungsvorschläge, ohne Dopplungen)	11	1

Tab. 2.
Ergebnisse der Reaktivität getrennt nach Versuchsbedingungen
*signifikant auf dem 5%-Niveau

damit sie bei den Verbalisierungen nicht unter Zeitdruck gerieten. Wie in Gruppe A (CTA) wurden auch die Probanden der Gruppe B (RTA) nach 5 Sekunden Schweigen zum Lauten Denken aufgefordert: „Was ist Ihnen hier durch den Kopf gegangen?“

4. Ergebnisse

Hypothese 1 besagte, dass der Erfassungsgrad (Sample Rate) der Blickbewegungen in Prozent bei Anwendung des CTA niedriger ausfällt als bei Anwendung des RTA. Die Auswertung zeigt einen leichten Trend zu einer höheren Sample Rate, wenn Probanden während der Aufgabendurchführung (und damit der Eye Tracking-Aufnahme) nicht laut denken. [Tab. 1] In der Tat haben 3 Personen aus Gruppe A (CTA) sich während der Durchführung dem Moderator zugewandt und ihm spontan Verbesserungsvorschläge vorgetragen oder Fragen gestellt, wozu sie nicht aufgefordert wurden. In dieser Zeit konnte der Eye Tracker keine Blickbewegung aufzeichnen, so dass die Sample Rate sank. Der Unterschied ist statistisch nicht signifikant.

Hypothese 4 (Probanden betrachten beim CTA mehr Bereiche (Areas of Interest)), kann anhand der hier vorliegenden Studie auch nicht eindeutig bestätigt werden, es lässt sich ebenfalls nur ein leichter Trend verzeichnen. Tatsächlich zeigen auch die Gaze Plots (Visualisierung aller Blickpunkte und Blicksprünge) aller Probanden aus Gruppe A (CTA), dass ihre Blicke nicht wie von den Autoren erwartet weit umherschweiften; diese konzentrierten sich vielmehr wie bei Gruppe B (RTA) auf den aufgabenrelevanten Bereich, d.h. der Suchmaske auf der Startseite. Der gefundene Unterschied muss als „zufällig entstanden“ betrachtet werden.

Eine statistisch bedeutsame Bestätigung konnte hingegen für die Hypothesen 2 und 3 gefunden werden: Probanden aus Gruppe A (CTA) haben innerhalb der ersten 45 Sekunden (= relevante Betrachtungszeit) deutlich mehr Fixationen produziert und ihre Gesamtfixationsdauer war wesentlich länger (p= .05).

Bis hierhin lässt sich feststellen, dass die Anwendung von CTA die Gesamtfixationsdauer und die Anzahl der Fixationen beeinflusst. Die Sample Rate und die Anzahl der betrachteten Bereiche (AOI) scheinen davon nicht betroffen zu sein.

Die letzen vier Hypothesen befassten sich mit dem Verhalten der Probanden in der Testsituation. Hypothese 5 besagte, dass sich die Aufgabendurchführung unter Anwendung des CTA verlängern würde. Aus Tabelle 2 wird ersichtlich, dass die Probanden beim CTA im Durchschnitt signifikant länger auf der Startseite verweilten (p= .05). Dies ist vermutlich den Verbalisierungen ihrer Handlungen, Gedanken und Eindrücke zuzuschreiben (Lautes Denken), die Zeit kosten. Hinzu kommt, dass 7 von 20 Probanden das Laute Denken während der Aufgabendurchführung z. T. sehr schwer fiel (im Gegensatz zu 4 von 20 Probanden in der Gruppe B (RTA)): Sie mussten während der Aufgabendurchführung bis zu 5 Mal vom Moderator zum Lauten Denken animiert werden („Was geht Ihnen gerade durch den Kopf?“), wobei die Probanden ihre Handlungen dann verlangsamt fortführten oder kurzzeitig stoppten, um etwas zu verbalisieren. Im Durchschnitt musste in beiden Gruppen jeder Proband einmal zum Lauten Denken aufgefordert werden. [Tab. 2]

Hypothese 6 behauptete, dass sich die Interaktion mit der Webseite zwischen

den Gruppen unterscheidet. In Gruppe A (CTA) wählten Probanden seltener einen alternativen Sucheinstieg, der sich als Link unterhalb der Suchmaske anbot: Nur 3 in Gruppe A (CTA) wählten diesen Link, während doppelt so viele aus Gruppe B (RTA) diesen Link klickten und sie somit auf die alten Suchprozessseiten führte. Dieses Ergebnis unterstützt das Phänomen der Reaktivität (Russo, Johnson & Stephens 1989), wonach CTA das Verhalten der Probanden beeinflussen kann.

Russo, Johnson und Stephens (1989) argumentierten auch, das Multi-Tasking von Lautem Denken und Aufgabendurchführung führe zu einer höheren kognitiven Belastung, die letztendlich zu häufigeren Bedienungs- und somit Usability-Fehlern führt. In der vorliegenden Studie wurden jedoch im RTA-Modus mehr Probleme gezählt. Dies gilt jedoch nur, wenn doppelt erfasste Probleme (über mehrere Probanden) berücksichtigt werden. Die Richtung ändert sich, wenn diese Doppelungen ignoriert werden. Hypothese 7 kann damit aufgrund fehlender Signifikanz nicht bestätigt werden und es kann auch kein genereller Trend in eine Richtung erkannt werden, da dieser je nach Auswertungsmodus umschlägt. In Anlehnung an vergleichbare Studien (z. B. van den Haak 2003) wurden die identifizierten Usability-Probleme in einem nächsten Schritt klassifiziert. **[Tab. 3]**

Die Klassifizierung der aufgetretenen Usability-Probleme zeigt vor allem den Zusammenhang zwischen einer längeren und intensiveren Betrachtung von relevanten Bereichen des Stimulus bei Anwendung des CTA (s. Ergebnisse zum Blickverhalten) und der Auffindbarkeit von Elementen: Die Probanden in Gruppe B (RTA) hatten häufiger das Problem, dass sie bestimmte Elemente auf der Webseite nicht finden konnten. 8 von 20 Probanden hatten mindestens ein Problem in der Kategorie „Layout", insgesamt wurden dazu 12 Probleme festgestellt (vgl. Tabelle 3).

Die letzte Hypothese lautete zugunsten des CTA: Es werden mehr spontane Vorschläge unter Anwendung des CTA geäußert. Während von allen 20 Probanden aus Gruppe B (RTA) nur eine einzige Person während des Gaze Replays einen Verbesserungsvorschlag äußerte („Es wäre zu überlegen, nach Warmmiete zu suchen, das wäre wesentlich einfacher, von der Suche her auch…"), so entstanden in der Gruppe A (CTA) 13 verschiedenste Verbesserungsideen (davon 11 ohne Doppelungen, vgl. Tabelle 2), die während der Aufgabendurchführung an den entsprechenden Stellen des Suchprozesses geäußert wurden. Diese reichten von einer neuen Sortierungsmöglichkeit der Suchergebnisse nach „Neueste Angebote zuerst" über „Darstellung der Suchergebnisse auf einer Karte" bis hin zum „Suchkriterium Warmmiete".

5. Diskussion

Signifikante Ergebnisse wurden für 3 von 8 Hypothesen gefunden: Bei Anwendung von CTA verlängert sich die Verweildauer und die Gesamtbetrachtungsdauer des Stimulus. Außerdem erhöht sich die Anzahl der erfassten Fixationen auf dem Stimulus. Das spricht dafür, dass die Methode des parallelen Lauten Denkens reaktiv ist, d. h. sich auf die genannten Parameter auswirkt. Aus diesem Grunde sollten Eye Tracking-Daten nur dann ausgewertet werden, wenn während der Aufgabendurchführung nicht laut gedacht wurde. Gerade bei der Feinanalyse (z. B.: „Wie viele Fixationen werden benötigt, bis der relevante Call-to-Action-Button gefunden und geklickt wird?") wären die mit Lautem Denken entstandenen Eye Tracking-Daten weder reliabel noch valide.

Die Sample Rate und die Anzahl der betrachteten Stimulus-Bereiche scheinen von der Reaktivität nicht beeinflusst zu sein. Ebenfalls nicht bedeutsam betroffen scheint die Interaktion mit dem Stimulus zu sein. Hier bleibt die ursprüngliche Aussagekraft der Daten auch unter Verwendung des parallelen Lauten Denkens erhalten.

Ebenso verhält es sich bei der Anzahl der identifizierten Probleme. Durch die Differenzierung dieser Probleme zeigt sich der Zusammenhang zwischen Anwendung des RTA und Problemen mit der Auffindbarkeit bestimmter Elemente, da hierbei besonders viele Layout-Probleme aufgedeckt werden. Dagegen wurden während des CTA deutlich mehr pro-aktive Verbesserungsvorschläge geäußert. Aus spontan vorgetragenen Verbesserungsvorschlägen lassen sich Schlüsse über das mentale Modell der Nutzer, fehlende Begeisterungs-Features, „missing links" u. Ä. ziehen. Auch können sie helfen,

	Gruppe A (CTA) N=20	Gruppe B (RTA) N=20
Terminologie (Begriffe auf der Webseite wurden nicht verstanden)	9	5
Dateneingabe (Probleme z. B. bei der Eingabe von Suchkriterien)	8	13
Feedback (nicht ausreichend oder erwartungskonform)	6	3
Navigation (Probleme z. B. mit dem Navigieren auf dem Immobilien-Portal)	3	4
Layout (gesuchte Elemente werden nicht gefunden)	1	12

Tab. 3.
Klassifizierung der aufgetretenen Usability-Probleme

grundlegende Nutzungsprobleme genauer zu identifizieren.

Durch Anwendung des CTA wird der Nutzer jedoch in eine unnatürlichere Nutzungssituation gebracht. Obwohl RTA mit Blickbewegungsmessung als Gedächtnisstütze und zur Aufmerksamkeitsmessung einen höheren technischen und zeitlichen Aufwand verursacht, ist dieser berechtigt, wenn ein natürlicheres Nutzerverhalten von Bedeutung ist.

Die vorliegende Studie zeigt, dass Probanden bei Anwendung des CTA weniger Probleme bei der Aufgabenbewältigung hatten, da sie mehr Zeit für das Explorieren und Auffinden von relevanten Elementen hatten. Hingegen können unter Verwendung des CTA wertvolle Hinweise von den Nutzern generiert werden. Es kann daraus geschlossen werden, dass RTA die natürlichere Nutzungssituation abbildet, CTA aber die „gründlichere Stimulus-Untersuchung" darstellt. Welche Methode vorteilhafter ist, hängt demnach davon ab, mit welchem Ziel eine Untersuchung durchgeführt wird (klassische Evaluation eines interaktiven Systems oder Generierung von neuen Features und Konzepten). Wenn dies feststeht, kann zwischen der Anwendung von RTA und CTA entschieden werden.

Die Reaktivität der CTA-Methode sollte weiter ansteigen, je freier eine Aufgabe im Usability-Test gestellt wird. Wenn Probanden z. B. gebeten werden, eine ihnen vollkommen neue Webseite zu explorieren und dorthin zu navigieren, wo sie möchten, so ist es wahrscheinlich, dass die Probanden mit CTA mehr Bereiche explorieren als diejenigen ohne. Diese Überprüfung wäre der nächste Schritt, um den genaueren Einfluss von CTA – auch auf das Blickverhalten bezogen – weiter zu erforschen.

Literatur

1. Ericsson, K. A. & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.
2. Nielsen, J. & Pernice, K. (2010). Eyetracking Web Usability. Amsterdam: Addison-Wesley Longman.
3. Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-259.
4. Russo, J. E., Johnson, E. J. & Stephens, D. L. (1989). The validity of verbal protocols. Memory & Cognition, 17 (6), 759-769.
5. Hyrskykari, A., Ovaska, S., Majaranta, P., Räihä, K.-J. & Lehtinen, M. (2008). Gaze Path Stimulation in Retrospective Think-Aloud. Journal of Eye Movement Research, 2 (4), 1-18.
6. Van den Haak, M. J., De Jong, M. D. T. & Schellens, P. J. (2003). Retrospective vs. Concurrent think-aloud protocols: Testing the usability of an online library catalogue. Behaviour & Information Technology, 22 (5), 339-351.