On the reliability of identifying design moves in protocol analysis

Gabriela Trindade Perry, Departamento de Design e Expressão Gráfica, Universidade Federal do Rio Grande do Sul, Av. Sarmento Leite, 320, sala 502, CEP 90040-060 Porto Alegre, Rio Grande do Sul, Brazil
Klaus Krippendorff, The Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA 6220, USA

This paper discusses issues and ways of measuring the reliability of segmenting verbal protocols of design activity, a central focus of design research. Reliability is an important issue in distinguishing 'design moves'. In the present study, seven students working for a master in design degree, one graduated designer and two professors segment a 30 min protocol of a product design process into design moves. The intra and inter reliability was calculated for these observers using alpha coefficients. Neither the students', designer's nor professors' segmentation reached the desired cut-off value of 0.8. This negative finding questions the clarity of existing conceptions and urges more concise definitions, better training of analysts, and formulating more decisive instructions.

Keywords: protocol analysis, design research, research methods

According to a bibliometric study (Chai & Xiao, 2011), protocol analyses, using the think-aloud method, detailed by Ericsson and Simon (1993), is one of the most popular design research methods. It involves distinguishing segments within a transcript of verbal accounts of design processes for further analysis. Regarding segmentation, Ericsson and Simon (1993, p. 205) state that 'the appropriate cues are pauses, intonation as well as syntactical markers'. Assuming that these criteria are objectively identifiable, Ericsson and Simon do not foresee reliability issues regarding this step of the analysis (p. 266). They do not mention non-syntactical criteria, which are important in design research and for which their assumption does not apply. Protocol analysis, as introduced in their famous book, aims at analyzing problems whose solving could be modelled – at least to some extent – by tools such as the problem behaviour graph (PBG), which would map the stage the problem solver is at, relative to a problem model. Examples of problems studied by Ericsson and Simon (1993) are: the tower of Hanoi, crypt arithmetic and theorem proving. Regarding design problems, Craig (2001) and Chi (1997) argue that protocol analysis using the think-aloud method might not be the most adequate method to analyze design processes.

Corresponding author:
Gabriela Trindade Perry
gabriela.perry@ufrgs.br

www.elsevier.com/locate/destud
0142-694X $ - see front matter *Design Studies* **34** (2013) 612–635
http://dx.doi.org/10.1016/j.destud.2013.02.001

Chi (1997) suggests modifications on Ericsson and Simon's method, for analyzing how representations change with learning and argues that, regarding ill-structured problems, it is not possible to know in which states of the problem space a problem solver could be. For this reason, she contrast her verbal analysis method with Ericsson and Simon's (1983) protocol analysis, proposing that verbal analysis should aim 'to capture the representation of knowledge that a learner has and how that representation changes with acquisition' (p. 3) — the emphasis on analyzing representations is strong on her paper. Chi states that her proposal differs from protocol analysis regarding (p. 4): 'the instruction, the goal or focus, the analysis, the validation, and the conclusion'. Chi (1997, p. 24) developed detailed instructions on how to segment verbal data, and emphasised the importance of measuring reliability of coded data in every step of the analytical process, i.e. 'during segmentation into units, categorizing or coding of the units, depicting the coded data, seeking pattern(s) in the depicted data, interpreting the pattern(s), and so forth'.

Several design studies used variants of the think-aloud technique and subsequently analyzed the protocol data. Gero and McNeill (1998), for example, developed a broad and wide coding scheme which was applied to 3 protocols of designing electronic devices, using the Delphi method. In their study, coders pause for ten days between each coding, which give the researchers the data needed to examine discrepancies between the two consecutive codings with the aim of locating disagreements and improving the coding scheme. The debate of emerging coding difficulties and revisions of problematic instructions constitute the 2nd and 3rd phases of the Delphi method were 'the group reaches an understanding of the issue' and 'disagreement is explored to bring out underlying reasons for differences and to evaluate them' (p. 34). Employing similar coding instructions as Gero and McNeill (1998), McNeill, Gero, and Warren (1998) investigated two hypotheses about the conceptualization and design of electronic devices. They also combined the Delphi method and a coding protocol, encouraging arbitration and the formation of consensus between coders. The authors argue that 'minimal disagreement between coders is desired but this can be difficult to achieve given the qualitative nature of the coding process' (p. 5). Recognizing the difficulty of achieving high reliability when analyzing textual matter, Krippendorff (2004, p. 3) argues that 'the mathematical complexity of analyzing variably unitized text, while an unquestionable hurdle for replicating research, is no justification for creating the methodological schism between quantitative and qualitative approaches to analyzing textual matter'. Whether the segmentation criteria are syntactical or conceptual, replicability of the coding process by independent coders is essential. Ball and Christensen (2009) also used protocol analysis of verbal data, but in a different way: they segmented the data using syntactical rules, trained one independent coder, and measured the inter-observer reliability [be-

tween that independent coder and the second author] using the Kappa coefficient. Another noteworthy study is by Carmel-Gilfilen and Portillo (2012), who explored differences in intellectual development between Architecture and Interior Design students. When classifying students, the authors report that trained raters coded students' data guided by a rating manual that includes examples, as well as agreement observed among raters. According to Campbell and Stanley (1963, p. 175) reliability, also called internal validity is 'the basic minimum without which any experiment would not be interpretable'. Campbell and Stanley use the term 'instrumentation' to describe threats to internal validity caused by differences in the way observers measure an event, for example (p. 179): 'if essays are being graded, standards may shift from event 1 to event 2'; 'if parents are being interviewed, the interviewer's familiarity with the parents may produce shifts [in the observation]'. Because reliability is a prerequisite for data to be interpretable, and can be measured only when coding is done by independent observers, this puts the burden of reliable data to the instrumentation or coding instructions, whether the method of recording data is quantitative or qualitative.

Apart from design research, two studies are noteworthy for their efforts to assure internal validity through instrumentation. The first is by Auld and White (1956), who developed a list of ten 'rules for unitizing', demonstrating that the segmentation of textual continua with non-syntactical or conceptual criteria is not a recent concern. The second comes from the field of linguistics. Carletta et al. (1997), present non-syntactical rules to identify and code moves in dialogues, reporting the reliability coefficients for segmenting [using kappa and percent agreement] and for categorizing the segments [using kappa coefficient only].

The present study uses the qualitative concept of 'design moves' as the criterion for unitizing/segmenting transcripts of design processes, relying on two alpha coefficients to assess their reliability. We believe design moves to be the essential ingredient of design processes and an important focus of design research.

1 Design moves as unitizing criterion in design research

'Design moves' can be defined as 'the smallest coherent operation detectable in design activity' and 'an act of reasoning that presents a coherent proposition pertaining to an entity that is being designed' (Goldschmidt, 1992). Goldschmidt (1995, p. 195) also characterizes design moves as 'a step, an act, an operation, which transforms the design situation relative to the state in which it was prior to that move.' She continues to say that 'moves are normally small steps, and it is not always easy to delimit a move in the think-aloud protocol of a single designer'. In Goldschmidt (1997, p. 447) yet another definition can be found: 'moves in the problem space are the small steps in which reasoning proceeds: i.e. they are representations of states and operators'.

Design moves have been used as a criterion for unitizing verbalizations in several design studies such as by: Kan and Gero (2008), Gero and Tang (2001), Cai, Do, and Zimring (2010) and Bilda and Demirkan (2003). Chai and Xiao (2011) also refer to Goldschmidt's early work (1991, 1995) on design process with several articles in their bibliometric study of the *Design Studies* journal, which points to a high acceptance of the concept of design moves and of protocol analysis in this field. The present study, it should be made clear, is not concerned with the concept, definition or utility of design moves, rather with the possible reliability of their identification.

Identifying design moves means locating their boundaries in a textual continuum — in the transcript of verbalized design processes. After their identification by observer/readers, analysts may, for example, search for links between moves based on references to previous moves. A network of relationship that could thereby be constructed is depicted as a linkography, which can be used to infer about the productivity of the design session, using measures such as: link index, back links/fore links and critical moves (Goldschmidt, 1992, 1995, 1997). In Figure 1, for example, design move n° 2 [highlighted in black] builds on design moves n° 1 and n° 0, and is geared to design move n° 4.

The rationale for using linkographies as an analysis tool is given by Van der Lugt (2001, p. 57): 'a well-generated idea can be expected to show signs of making use of the information gained earlier in the process'. According to Goldschmidt and Weil (1998, p. 90), 'a link between two moves is established when the two moves pertain to the same, or closely related, subject matter'. Since these links are defined by the observer/coder's 'common sense', Van der Lugt (2001, p. 61) expresses concerns about the reliability of establishing such a network, stating that 'this way of coding is highly open to subjectivity'.

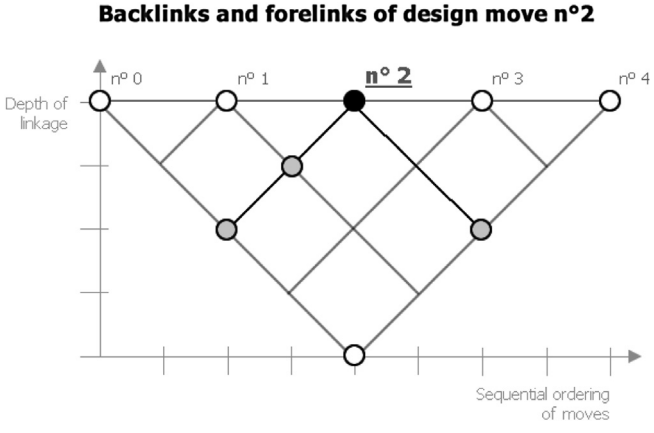

Figure 1 An example of a linkography. Each numbered circle represents a design move

There are studies which used sophisticated mathematical tools to analyze linkographies, such as Kan and Gero (2008), who used the x and y coordinates of the linking network of linkographies as the input for a cluster analysis to infer the underlying design activity. Cai et al. (2010) also used the x and y coordinates of linkographies as input for distance graphs to infer sources of inspiration in the design activity.

Bilda, Gero, and Purcell (2006) and Van der Lugt (2001) used linkographies to draw inferences about the effects of sketching in conceptual design. The authors of these studies were concerned with the reliability of assigning links to design moves: they searched for objective, direct reference using verbalizations, gestures and drawings as traces of references. Bilda et al. (2006) inspected designer's intentions complementing verbal data, with video recordings and drawings: their actions were used as clues to find changes in intentions (p. 592). They verified the content of each segment and inspected related segments to see if there was a connection. For the cases of distant links [in the timeline], a strategy based on a list of frequent words was developed. Van der Lugt (2001) used 'design ideas', a concept similar to design moves. To find the links between designers' intentions, he developed five guidelines for observers to identify links between 'design ideas'. We are not aware of a quantitative approach to assess the reliability of these connections.

We need to note that most of the cases mentioned above focused not on the segmentation of transcripts but on the assignment of categories to the identified segments. This choice rendered the assessment of the reliability of segmentation of minor importance. One could speculate that the omission of concerns for the reliability of unitizing/segmenting is due to the absence of ways of measuring it.

Especially in the construction of linkographs, we believe that assessing the reliability of segmenting protocols into design moves is central for establishing trust in their construction. If the segmentation of protocols of design activities is unreliable, a linkograph built on them needs to be questioned. This is not to say that the reliability of segmentation is sufficient. It merely is a necessary first step. If the segmentation is unreliable, it is likely that they give rise to conflicting linkographies. Even if observers perfectly agree on their independent segmentation, it is not unthinkable that two analysts would develop unlike linkographies from the same stream of design moves. This is to say that we ultimately need to assess the reliability of linkograph constructions as well, but this goes beyond the more modest aim of this paper. We will rely on the alpha coefficients that are outlined in the next sections of this paper.

2 *Agreement coefficients for unitizing data*

If readers cannot agree on where in a protocol of design activities a design move begins and where it ends, conclusions drawn from an analysis of these moves need to be questioned. High inter-observer agreement is necessary to assure the interpretability of the recorded data. Their reliability can be inferred from agreements that have been observed under carefully controlled conditions. Krippendorff (2011, p. 1) defined reliability as 'the extent to which different methods, research results, or people arrive at the same interpretations or facts' and proposed several agreement coefficients for content analyses and other inquiries that make use of textual matter. He points out that reliability is a necessary but not sufficient condition for validity [which is beyond the scope of the present study].

Krippendorff's (2013) family of alpha coefficients applies to data that are generated by several analysts, coders, observers, or in the case of texts, readers, who attend to the same set of phenomena – the protocol of a given length – working independently from each other and following the same instructions. Working independently from each other is important as collaboration would invalidate the measured agreement. Carefully worded instructions are also necessary as they are the only way to link the resulting data to the phenomena of interest. In its general form, alpha is defined by:

$$\alpha = 1 - \frac{D_o}{D_e}$$

D_o is a measure of the observed disagreement among methods, results, or observers. D_e is a measure of the expected disagreement, the disagreement that would be observed if the data were chance events. The latter could result from the failure of carefully examining the phenomena to be recorded, by providing ambiguous instructions, or messing up the data. From its algebraic form, one can see that $\alpha = 1$ when $D_o = 0$, indicating the condition of perfect agreement or perfect reliability. $\alpha = 0$ when $D_o = D_e$, which would indicate that the data making task equals chance, and that the data have no relationship to the phenomena of analytical interest. α can be negative if the observed disagreement exceeds expectations. Because perfect reliability is difficult to achieve, social scientists commonly require $\alpha \geq 0.8$ for data to be taken seriously. Should α be lower than that, $0.8 > \alpha > 0.666$, data may be used for cautious explorations but not for drawing firm conclusions (Krippendorff, 2013, p. 325).

The rationale of the alpha coefficients for unitization, which are used in this paper, is found in Krippendorff (2004, p. 8): 'for reliability to be perfect, the units that different observers identify must occupy the same locations in the continuum and be assigned to identical categories. Disagreements sum deviations from this ideal by counting the pair wise differences between units and

gaps, one pair at a time. Intuitively, such differences must be zero when units perfectly coincide. They must increase as the overlap between any two units lessens and reach their largest value when a unit does not overlap with any other unit.'

The present study concerns the identification of design moves in protocols, i.e., units or segments of text in a continuum. We chose design moves because we consider them to be a valuable tool for investigating the design process. Because they are not a syntactical criterion, Goldschmidt (1995, p. 195) points out that 'it is not always easy to delimit a move in the think-aloud protocol of a single designer'. If we want to say something about design moves, an assessment of reliability of the observers – who actually segment the transcripts – is essential. While in content analyses, reliability assessments are standard requirements, Krippendorff (1995, 2013) observed that they mainly address the reliability of coding predefined units. Perhaps for lack of a simple agreement measure for unitizing, the reliability of unitizing is often ignored although segmenting can be unreliable as well. Artstein and Poesio (2008, p. 580) report similar limitations in the field of Computational Linguistics. There, it is 'the practice to assume that the units are linguistic constituents (that) can be easily identified, such as words, utterances, or noun phrases, and therefore there is no need to check the reliability of this process' – just as Ericsson and Simon (1993) advised. Artstein and Poesio (2008) discuss several known reliability measurements but point to Krippendorff's (1995) α_U as the most adequate coefficient for unitization. They also mention that, to their knowledge, α_U has never been applied. The only reference found using the α_U coefficient was Yalçinkaya (2010), who used it to develop a software tool. As the issue of the concepts behind the several ways to measure reliability of segments and the advantages/disadvantages of each one is out of the scope of this paper, we refer the interested reader to Artstein and Poesio (2008) and Krippendorff (2013).

Meanwhile, two developments have come to our attention. In the 3rd edition of his content analysis text, Krippendorff (2013, p. 309–315) modified his 1995 α_U to embrace the coding of units into categories, now called ${}_u\alpha$, and in personal communication, Krippendorff developed an α coefficient for distinctions, called ${}_d\alpha$. As ${}_d\alpha$ is unpublished at the time of writing the present paper, we encourage interest readers to contact Krippendorff (2012).

In the present study, we use the 1995 α_U coefficient for unitizing a continuum, accommodating omissions of irrelevant matter without requiring segments to be coded, and the ${}_d\alpha$ coefficient for drawing distinctions within a continuum regardless of whether the resulting segments are relevant or irrelevant and whether they are coded. Although α_U is more appropriate for our protocol analysis, we chose to also report the ${}_d\alpha$ values of for two reasons: to compare the two coefficients – with or without omitted matter, and to introduce this

new coefficient to the research community. The first author developed open software to calculate all three coefficients. Interested readers may download it from http://www.gabriela.trindade.nom.br/2013/02/calculating-alpha-d-and-alpha-u/. Data formatting instructions are provided on the interface of this software. The link also provides the data files used in this study as examples. Since some of their computations are complex, the following merely defines the difference functions underlying these three coefficients and verbalizes the observed and expected disagreements. Interested readers are referred to the original publications (Krippendorff, 1995, 2004a, 2012, 2013).

2.1 The α_U coefficient, as published in 1995

As already mentioned, the α_U coefficient measures the agreement among any number of observers who unitize a continuum, accommodating the omission of irrelevant matter without requiring relevant segments to be coded. It responds to differences in all paired segments of relevant matter. We graphically exemplify the difference function of α_U between three segments in two unequally segmented continua. Bold lines representing relevant matter (Figure 2):

The observed disagreement ${}_UD_o$ is defined as the average difference ${}_U\delta^2$ between all pairs of overlapping segments in the continuum. By contrast, the expected disagreement ${}_UD_e$ is the average of all combinatorially possible differences.

2.2 The ${}_d\alpha$ coefficient, as developed in 2012

Krippendorff's (2012) ${}_d\alpha$ coefficient measures the agreement of distinctions introduced by any numbers of observers in a continuum regardless of whether the resulting segments are irrelevant or relevant matter and regardless of whether they are assigned to categories or are valued. For how Goldschmidt (1991, 1992, 1995, 1997) identified design moves, ${}_d\alpha$ would have been the most appropriate reliability measurement as she unitized the whole protocol without exception and without categorizing the identified design moves. However, unlike Goldschmidt's approach, we allowed observers to omit irrelevant segments of the protocol, which rendered α_U, not ${}_d\alpha$, the preferred coefficient.

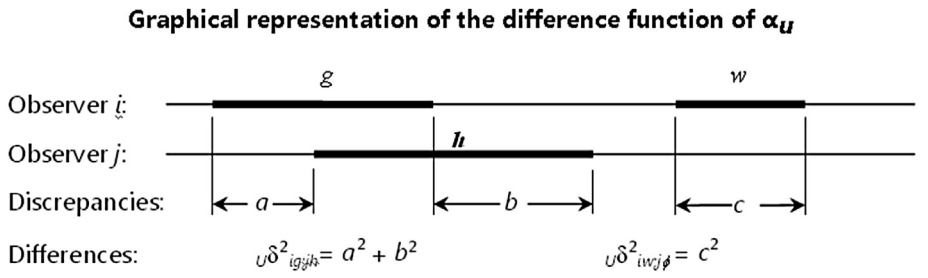

Figure 2 Difference function of α_U, taking included and excluded segment into consideration

Because $_d\alpha$ is attractive for its simplicity and useful for analysts whose data are similar to that of Goldschmidt, we decided to obtain $_d\alpha$ for comparison.

To appreciate the differences between any two distinctions to which $_d\alpha$ responds, the following graph compares one distinction g made by observer i with two distinctions h and $h-1$ made by observer j (Figure 3):

In other words, if one observer's distinction falls within the interval (segment) between two distinctions made by another observer, the difference is the square of the smallest discrepancy. Much as for $_u\alpha$, the observed disagreement $_dD_o$ is the average difference $_d\delta^2$, for all distinctions by one observer paired with those of another; and the expected disagreement $_dD_e$ is the average of all combinatorially possible differences.

2.3 The $_u\alpha$ coefficient, as published in 2013

Krippendorff's (2013, p. 309–315) $_u\alpha$ coefficient responds to the assignment of categories to the identified segments. It assumes that observers who distinguish two adjacent segments must have a conceptual reason for it, which is expressed in assigning unlike codes to them. The following graph illustrating the difference between pairs of unlike valued segments may be compared with the one presented for the agreement coefficient $_u\alpha$ (Figure 4).

Just as for α_U, the observed disagreement $_uD_o$ is the average difference between all pairs of overlapping segments; and the expected disagreement $_uD_e$ is the average of all combinatorially possible differences among segments regardless of their original position in the continuum.

We are mentioning this coefficient for comparison with what α_U and $_d\alpha$ responds to. As already mentioned, in the present study segments were not categorized and $_u\alpha$ therefore was not applicable.

3 Methodology

The steps towards generating data to assess the reliability of identifying segments were: (1) prepare and run a design assignment; (2) record and transcribe

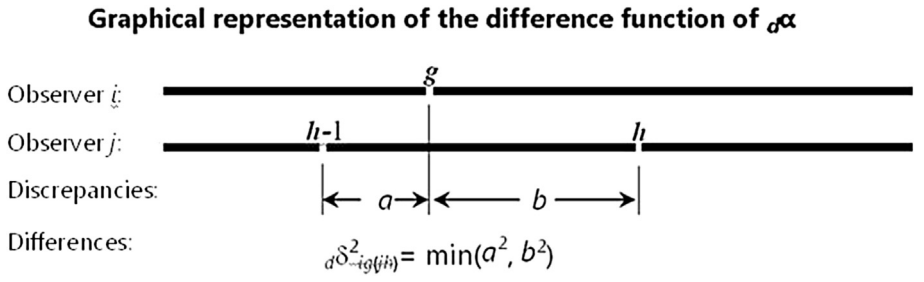

Figure 3 Difference function of $_d\alpha$: all segments are considered as included

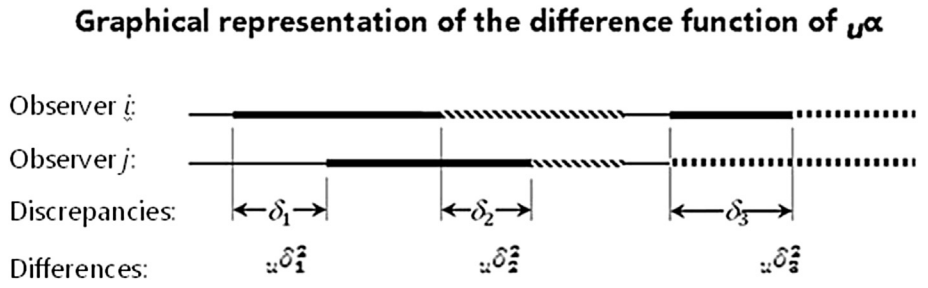

Figure 4 Difference function of $_u\alpha$: Considering excluded and included segment, with their categories

what was said; and (3) prepare the transcripts for comparison across observers who segmented the protocol. Eleven students, two designers and two professors volunteered in the experiment. We expected that the three groups would differ in their segmentations of the transcript. It is reasonable to assume that: students would be less committed and knowledgeable about what constitutes design moves than the professors, and do worse than the professors. We expected that the designer would do best as they made the moves that the students and professors were asked to identify.

3.1 The design assignment

One designer, graduated in 2009, with three years of experience in furniture and product design finishing his master in design in 2012 volunteered for the task. He was supposed to design the furniture for the 'nap room' in a kindergarten. He was given the design brief and was instructed to say out loud whatever came to mind while designing the furniture. However, he reported being uncomfortable with the procedure of thinking-out-loud, as he said it 'impaired his ability to think'. Considering this impediment, we asked whether he would accept taking part in a different design assignment, working with a second designer. Instead of having to think-out-loud alone, he was asked to explain his thoughts to that second designer, which is a more natural assignment than talking in monologue to a tape recorder. The second designer had been a colleague of the first in his graduate studies, and works in the same field (although not in the same company). He finished his master in production engineering in 2011. Both reported being very comfortable with the experiment setup — one of them said that he 'even forgot it was being recorded,' and that 'the exercise was very fruitful and pleasant, and that maybe we [they] should try it more often'.

The assignment was to design the check-out counter for a department store. It was based on Guimarães, Diniz, and Silva (2002), who reported the design constraints to design this counter. The choice to use an existing design case was motivated by the convenience of providing the designers with real data about a problem and its environment. This design session lasted for 2 h, and was recorded with a full HD camera, which provided good sound and image

quality. Since the video data were too lengthy for the purposes of the segmentation exercise, the decision was to use the first 30 min only.

From our view point, and in full agreement with the designer, the method of generating think-aloud protocols is not as natural to what is going on in design processes as is explaining, arguing, and collaborating with someone else. Dialogue seems more appropriate than monologue, as our designers confirmed. We agree with Craig (2001) in that it seems very difficult to infer the underlying cognitive processes from think-aloud data, especially when problems are ill-structured and verbalizations are not readily at hand. We need to note that Ericsson and Simon (1993) applied the think-aloud method for generating protocols to more well structured problems – problems whose solution stage could be more easily articulated and sequentially traced. In this sense, Ericsson and Simon (1993) had more data than only the protocols, as they had a model of the problem space which could be used to trace the problem solver's path towards a solution. Perhaps the term 'protocol' and 'think-aloud' distracts us from acknowledging the social situation in which humans talk and reveal their thoughts to each other and to the analyst as well.

Generating verbal accounts of design processes with more than one subject, although not the rule in design research, is not uncommon. Observations of teamwork dates back to the 1980s and 1990s (e.g. Valkenburg & Dorst, 1998; Cross & Cross, 1995). Austin (2001) designed a fairly similar setting to investigate multidisciplinary teams of five designers, whose assignment was related to architecture. Ball, Onarheim, and Christensen (2010), Tang, Aleti, Burge, and van Vliet (2010) and Christiaans and Almendra (2010) analyzed the design process of three pairs of software engineers designing educational software for traffic control.

3.2 *The observers and the segmentation assignment*

The transcription resulted in an eleven page document, containing 3530 words. It was edited to show the time each designer spoke in the far left column, the transcriptions in the centre column and references to drawings in the far right column, everything aligned with the text.

The observers who undertook the segmentation of this document consisted of eleven students (masters in design), who were studying protocol analysis in design research for the first time, two recently graduated designers, and two professors who were the teachers of the eleven students. They were presented seven slides of definitions taken from the four Goldschmidt papers: 1992, 1995, 1997 and 1998. Participants were given the opportunity to discuss these definitions and relate them to the examples. After being so instructed, all students assured the experimenter that they understood the instructions, which were shown in English – not their first language.

3.3 Collecting data: the segmentation task

Two segmentation sessions (3 h each) were conducted within a one week interval. The eleven students and two designers were asked to bring a headset not to interfere with each other while watching and listening to the video. Before starting the task, they were given a warm-up exercise, which consisted in segmenting a two page transcript of an unrelated design assignment. After finishing, each student returned the printed transcript with the annotations to the experimenter.

Right after starting the 1st session, there surfaced several uncertainties about the segmentation procedure, mostly regarding the identification of boundaries of design moves. The experimenter pointed out that those questions would not be answered because it would introduce a bias in measuring agreement. There were also questions regarding: the need to identify arguments (Goldschmidt, 1992); whether it was possible to exclude text snippets; whether it was possible to have overlapping moves; what to do in case of repetitions (whether they should be marked as the same move), and how to indicate the start and end of moves. After the session, all questions were answered for everyone to hear.

One week after the 1st session, the 2nd session was carried out. In response to questions raised after that session, the procedure was repeated with two modifications. One was the introduction of a clear way to indicate the start and the end of design moves: it was agreed that everyone would use brackets, excluding segment outside brackets and preventing overlapping segments to occur. The second modification was the introduction of an exercise for identifying move boundaries based on protocol excerpts from the four Goldschmidt papers. Six slides were added to the presentation shown in the 1st session [which had only the definitions of design moves], containing protocol snippets from each of Goldschmidt's four papers. The students were asked to identify the design moves in those snippets. After that, the professors drew brackets over the text [projected on a wall], identifying design moves as they were indicated in each paper, for example:

> *[If I look at the form again, it seems that spatially, these are the larger directions] [I am getting one, two, three spaces here and one, two there] [They're about square, so there is a tendency to try and see them as spaces] [These are secondary directions within the space, so the entry is actually moving in along the secondary directions]*

It is important to note that Goldschmidt's examples did not exclude irrelevant segments. After these clarifying modifications, the students and the designers received new copies of the same transcript they unitized in the 1st session.

These adjustments, from the 1st to the 2nd session, could have affected intra-rated agreement, as the two unitizing experiments were not the same. We were

aware of this when the changes were made. However, we decided to make these changes because the identification of design moves was not clear to the students, and we wanted to know whether these changes would improve their performance.

Two students did not show up for the 2nd session and two students could not finish the task [one had to leave before finishing the task]. One of the designers marked overlapping moves in the 1st session, which cannot be treated properly by the agreement coefficients. For this reason, we could not use his data in this study.

The professors segmented the transcript for the 1st the week after the 2nd segmentation session, and segmented it for the 2nd time a week later. The complete data collecting procedure consisted of four segmentation session and lasted four weeks.

3.4 Preparing the data

Data from four students and one designer had to be discarded for the reasons mentioned above. A visual inspection of the remaining data shows that the criteria used for demarcation varied considerably: some students had not marked a single move in a whole page, while others were more generous in their identification. Also, because of the changes introduced in the 2nd session [the way boundaries were to be indicated and their exemplification by Goldschmidt's examples], some students who had not excluded any segment in the 1st session excluded segments in the 2nd. A total of 18 data sets were analyzed: 7•2 from the students; 1•2 from the designer and 2•2 from the professors. These data were manually formatted as cvs files and used as input to the software that calculates the α_U and ${}_d\alpha$ coefficients, mentioned earlier.

4 Results

The intention was to compare students', designers' and professors' segmentations. However, since one of the designers marked overlapping moves, that material had to be discarded. Only intra-observer agreement could be calculated for the designer.

4.1 Agreement of students' segmenting

The values of the α_U and ${}_d\alpha$ coefficient for students' inter-observer agreement on the 1st session, taken pair wise, are shown in Tables 1 and 2.

Evidently, no pair of students reached the $\alpha_U = 0.8$ target value. This finding is due to too many discrepancies between the identified design moves. The ${}_d\alpha$ was not so severe, since it did not respond to the disagreements between included and excluded segments. The pair St.07 + St.05 reached ${}_d\alpha = 0.87$ − meaning they made the segmentation using the same criterion. However, their α_U value was close to zero, indicating chance agreement. Both students had about ½ of

Table 1 α_U **coefficient, inter-observer agreement for the 1st session**

	St.01	*St.02*	*St.03*	*St.04*	*St.05*	*St.06*	*St.07*
St.01	–	−0.22	−0.35	−0.01	−0.57	0.05	0.28
St.02		–	0.09	0.21	0.27	0.27	0.09
St.03			–	0.31	0	0.20	0.07
St.04				–	0.18	0.33	0.12
St.05					–	0.13	−0.02
St.06						–	0.26
St.07							–

their segments excluded, which contributes to a high disagreement in α_U. Figure 5 shows an excerpt of the segmentation made by these two students. This image was generated by the above mentioned software developed for this study. The image is not complete, because the segmented text is too long to be shown here, so we suggest the reader run the software to see the complete image.

Evidently, α_U severely penalizes two overlapping segments with different values, and inspecting Figure 5 it is possible to note that students 05 and 07 disagree a lot in this matter. This is the reason for ${}_d\alpha$ to be higher than α_U. In conclusion, if the coefficient of choice was ${}_d\alpha$, the differences in identifying boundaries would be acceptable, meaning these students used the same criteria to identify design moves. However, when excluded segments are taken into account, the differences get large enough to claim that these students did not use the same criteria to identify and categorize design moves. The results for the 2nd session are shown in Tables 3 and 4.

In the 2nd session, no pair of students reached the target value, neither for α_U nor ${}_d\alpha$.

To test the hypothesis of difference between the inter-observer agreement with both coefficients the 1st and 2nd sessions, we ran two Wilcoxon signed ranks tests. The results point to no difference at the 0.05 level between sessions for α_U: $W(21)$, $Z = -1.48$, $p = 0.14$.

Table 2 ${}_d\alpha$ **coefficient, inter-observer agreement for the 1st session**

	St.01	*St.02*	*St.03*	*St.04*	*St.05*	*St.06*	*St.07*
St.01	–	−0.22	−0.33	−0.14	0.47	−0.04	0.5
St.02		–	0.22	0.43	0.43	0.63	0.35
St.03			–	0.5	0.31	0.27	0.5
St.04				–	0.18	0.33	0.16
St.05					–	0.5	0.87
St.06						–	0.56
St.07							–

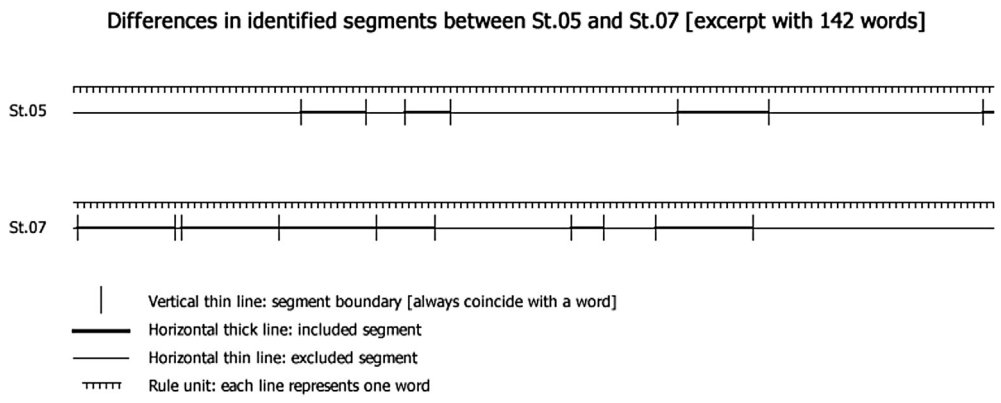

Figure 5 Unitizations made by St.05 and St.07 in the 1st session

Regarding ${}_d\alpha$, the Wilcoxon test pointed to significant differences between the sessions: $W(21)$, $Z = -3$, $p = 0$. This result points to a negative effect from the training. We did not expect this result, as we introduced the new procedures [a graphical convention to move boundaries and an exercise on identifying moves from Goldschmidt's paper] with the intent of improving inter-observer values. We decided to use a non-parametrical test because a Shapiro–Wilk, at the 0.05 level, pointed to the non-normality of α_U, ${}_d\alpha$ and of the amount of identified segments [respectively $p = 0.03$, $p = 0.02$ and $p = 0$]. It was also noted that number of excluded segments increased in the 2nd session. In the 1st session, four students [out of seven] marked all units for inclusion, while only one did so in the 2nd session. The hypothesis of the difference in the number of excluded segments between sessions was tested, using a Wilcoxon test. The ratio between the excluded segments and the total number of segments identified per student was computed [to take the total number of segments into consideration]. There was a significant difference [$p = 0.04$] at the 0.05 level, meaning there were more excluded segments in the 2nd session. We consider that it is likely that the introduction of a graphical convention to mark segment boundaries could be the cause of this result, because, in the 1st session, some students had used the '|' sign to mark boundaries, rendering the task of identifying excluded segments impossible. We consider it is not likely that this increase in

Table 3 α_U coefficient, inter-observer agreement for the 2nd session

	St.01	*St.02*	*St.03*	*St.04*	*St.05*	*St.06*	*St.07*
St.01	–	−0.38	−0.23	−0.32	−0.66	0.2	0.23
St.02		–	0.44	0.52	−0.22	0.16	0
St.03			–	0.36	−0.22	0.42	0.23
St.04				–	−0.4	0.26	0.03
St.05					–	−0.2	−0.8
St.06						–	−0.1
St.07							–

Table 4 $_d\alpha$ coefficient, inter-observer agreement for the 2nd session

	St.01	*St.02*	*St.03*	*St.04*	*St.05*	*St.06*	*St.07*
St.01	–	−0.5	−0.44	−0.42	−0.56	−0.41	0.18
St.02		–	0.38	0.61	0.17	0.39	0
St.03			–	0.54	0.14	0.5	0.16
St.04				–	−0.27	0.6	0.2
St.05					–	−0.06	−0.55
St.06						–	−0.27
St.07							–

the amount of excluded segments could have been caused by the exercises of identification of design moves based on Goldschmidt's papers, as all her examples show only included moves. However, it is important to be reminded that the excluded segments have no impact on $_d\alpha$ [the coefficient negatively affected by the training].

Regarding the exercise of identification of design moves, it might have brought confusion to the students who were trying to understand how to identify design moves – a task they reported being very hard. To dispel these doubts, another round of experiments shall be performed, with new subjects and with a detailed instructional material. Measuring inter-observer agreement is important, since replicability requires observers to work independently of each other and be, hence, interchangeable. If inter-observer agreement is high, there is evidence to claim that the data generated do represent phenomena that have been seen alike by observers, that the instructions are clear enough so that other trained and independently segmenting observers would be able to reproduce the data. Intra-observer agreement, on the other hand, is important because it measures an observer's 'stability' or 'consistency,' i.e. the ability to achieve similar results every time he/she applies the instructions to segment a text continuum. Although intra-observer agreement does not say much about the reliability of the data, it can be important to locate the source of unreliability in the observers. For this reason, values of intra-observer agreement are shown in Table 5. Again, no observer reached the target 0.8 value.

The intra-observer agreement for student St.05 was higher than the others, demonstrating she was more stable while segmenting the data. Students St.04, St.03 and St.06 also had a better performance than the other students, whose intra-observer agreement coefficients were close to chance. None of

Table 5 α_U and $_d\alpha$ coefficients, intra-observer agreement

	St.01	*St.02*	*St.03*	*St.04*	*St.05*	*St.06*	*St.07*
Results for α_U	0.45	0.09	0.44	0.45	0.69	0.34	−0.7
Results for $_d\alpha$	−0.02	0.37	0.56	0.64	0.77	0.45	−0.64

these students had prior training and experience in the protocol analysis method. The design of this experiment – purely quantitative – does not allow any assumptions about the reason these four students performed better – although not well enough to reach the 0.8 target value. We cannot tell for example, whether these students differed in motivation, carefulness, or understanding the instructions. In further studies, we shall add an interview with the observers, to trace reasons for these differences in performance. However, although these four students were more stable than the others, they did not share the same procedure to segment the data. Inspecting Tables 1–4, it is possible to see that no pair of these students shows a better performance [inter-observer agreement] than the others. The Mann–Whitney test for the difference of performance [values of α_U and ${}_d\alpha$] regarding students St.03 + St.04; St.03 + St.05; St.03 + St.06; St.04 + St.05; St.04 + St.06 and St.05 + St.06 versus all other possible pairs indicated no significant difference, at the 0.05 level. Regarding α_U, the p values were $p = 0.12$ for the 1st session and $p = 0.6$ for the 2nd, a case in which the null hypothesis of no difference between the mean ranks of these pairs cannot be rejected. Regarding ${}_d\alpha$, the p values were $p = 1$ for the 1st session and $p = 0.15$ for the 2nd, leading to the same conclusion.

Therefore, we conclude that these four students were stable at their segmentation, but each one was developing his/her own understanding about how to segment verbal data using design moves – which is not a desirable result. We would want a training course to assure that the observers would achieve a high inter-observer agreement, meaning they all are using the same rules to segment the data.

4.2 Intra-observer agreement of designer's segmenting

The intra-observer agreement values of the α_U and ${}_d\alpha$ coefficients for the designer are presented in Table 6. It was expected that he would reach a higher value than the students, since he would know how to identify the design moves he made. That was not the case, as there were students who achieved similar results (see Table 5, with students' intra-observer agreement results). Once again, ${}_d\alpha$ is less severe than α_U, for the reasons discussed before. The reason the values of both coefficients is so different for the designer is that he marked no segments as excluded in the 1st session and several units as excluded in the 2nd session – α_d does not recognize the difference between included and excluded segments, hence the difference.

Table 6 α_U and ${}_d\alpha$ coefficient, intra-observer agreement

	Des.01_01 & Des.01_02
Results for α_U	0.32
Results for ${}_d\alpha$	0.6

4.3 Agreement of professor's segmenting using α_U and $_d\alpha$

When it comes to the professors, the inter-observer values from the 1st and 2nd session increased for both coefficients, which would point to a positive effect from training. However, since there were few subjects, it is not possible to draw statistical conclusions about the difference between values shown in Tables 7 and 8. It was expected that they would have the highest agreement comparison with the students, since they were [probably] more motivated and committed to the research. However, they also did not reach the target 0.8 value.

For reasons explained earlier, the intra-observer values for both coefficients are presented. As expected, the professors were more stable coders. This was manifest in the fact that their excluded segments had similar lengths and were in similar positions on the continuum in both sessions, thus α_U is not much more severe.

5 Conclusions

The segmentation of protocols into design moves was studied with the help of two agreement coefficients: $_d\alpha$ and α_U. For data with excluded segments, α_U is advised. If only the agreement regarding distinctions [boundaries of two segments] is of interest, then $_d\alpha$ is recommended. This is due to the unequal responsiveness of the two coefficients, but also born out in our study.

Regarding the data collected, only the students' data were numerous enough to allow for statistical comparisons. In the most sensitive case – the difference in students' performance as observers in the 1st and 2nd sessions – a Wilcoxon signed rank test pointed to no significant difference at the 0.05 level in the case of α_U, but did point to a difference regarding $_d\alpha$. It means that, when it comes to identify boundaries, training had a negative effect on students. In the case of $_d\alpha$ students performed better in the 1st session, before the introduction of a graphical convention to mark boundaries and of exercises of identification of design moves, taken from Goldschmidt's papers. These changes were introduced, in the 2nd session, with the intention of improving students' understanding of the concept of design moves. However, it did not happen. As we did not interview the students after each session, we cannot tell the cause of this counterintuitive effect.

Table 7 α_U and $_d\alpha$ coefficient inter-observer agreement

	Prof.01_01 & Prof.02_01	*Prof.01_02 & Prof.02_02*
Results for α_U	0.58	0.65
Results for $_d\alpha$	0.58	0.7

Table 8 α_U and ${}_d\alpha$ coefficient intra-observer agreement

	Prof.01	*Prof.02*
Results for α_U	0.63	0.56
Results for ${}_d\alpha$	0.58	0.63

Regarding the amount of excluded segments, a Wilcoxon test pointed to a significant difference in the ratio between the number of identified segments and excluded segments per student [$p = 0.04$]. However, it is important to remind that the excluded segments have no impact on ${}_d\alpha$ [the coefficient negatively affected by the training].

Despite of these findings, we cannot help noticing that the effect of training was doubtful. As said before, the changes introduced in the 2nd session did not improve students' performance. All of the students were graduated as designers or architects, and, although not all of them have experience designing furniture, they all have experience as professionals at their area [between three to five years prior to entering the post-graduation program]. Five of them are also design teachers in graduation level courses. We had assured that the students had read Goldschmidt's papers, as they had been discussed them in classroom earlier. The students had been reading and discussing related literature for two and half months, at that point. Because none of the students had personal interest in protocol analysis and the analysis of design moves, we speculate that none of them was particularly motivated to excel in this experiment. We sampled these students because they had appropriate knowledge in design and analytical skills, and assumed that they were at least curious, but their performance was disappointing. The value of the α coefficient for all 7 students and the 2 professors are summarized in Table 9.

Professors' agreements were much higher than students', considering both coefficients. This might be the consequence of: prior learning, while students were training; using a 'stable' training material and being committed to the research. However, there was one student who was a more 'stable' coder than both professors, as Table 10 shows.

Student St.05 had the best overall performance as observer/coder, with intra-observer agreement of α_U equal to 0.69. She was followed by the two

Table 9 Results of inter-observer agreement of students and professors

	All students' 1st session	*All students' 2nd session*	*Professors' 1st session*	*Professors' 2nd session*
Results for α_U	0.08	−0.05	0.58	0.65
Results for ${}_d\alpha$	0.44	0.03	0.58	0.7

Table 10 Intra-observer agreement, students, professors and the designer

	St.01	*St.02*	*St.03*	*St.04*	*St.05*	*St.06*	*St.07*	*Des.01*	*Prof.01*	*Prof.02*
Results for α_U	0.45	0.09	0.44	0.45	0.69	0.34	−0.7	0.32	0.63	0.56
Results for ${}_d\alpha$	−0.02	0.37	0.56	0.64	0.77	0.45	−0.64	0.6	0.58	0.63

professors. The designer did not have a better intra-observer agreement than students St.03, St.04 and St.06.

In content analysis – as well as in many design studies – identifying boundaries and marking segments as excluded/included is only the first step of generating analyzable data. Usually, after boundaries are identified and irrelevant matter is distinguished from relevant matter, observers assign categories to these segments, and conclusions are based on the frequency of these categories. In these cases, ${}_u\alpha$, as published in 2013 and discussed in part 2.3 of this paper would be the correct choice. In the present study, however, the focus was on drawing distinctions and identifying relevant matter. The rationale is that identifying design moves is the 1st step for drawing a linkography, a widely used representation in design research, from which the researcher can infer the productivity or creativity of the observed designer. In this view, assessing the reliability of segmentation would not be sufficient, but a necessary step before drawing a potentially trustworthy linkography. Lacking suitable measures for obtaining the reliability of connecting different design moves into a linkography, the reliability of segmentation is all we have right now.

This study also points to the importance of training observers to be able to reliably segment textual protocols into distinct design moves. Because the criterion for deciding what a design move is – and defining where it starts and ends – is largely conceptual, much work needs to be done to clarify the 'design move' concept and translate it into reliable segmentation instructions. Only then is it possible to infer the intention of the designer. As expected, observers with little training exhibited poor agreements and generated data that could not be used to report research results. We wish to stress that it was not the aim of this study to question 'design moves' as a good concept for guiding design research. We chose to focus on it because it is widely used, and we deem it instrumental in investigations of design processes.

Concerning future directions, we think the following would improve the analysis of protocols for the design moves they manifest:

- Generating the protocol of design activity in dialogue has proven to have distinct advantages over the monological think-aloud method, but it presents analytical challenges that the think-aloud method does not face.

Key among them is who, if any one, leads the process. We believe much can be learned from the transcription practices of conversation analysis.

- Developing clearer definitions of design moves and more concise instructions for how they can be identified in transcripts of design processes is essential for any progress. Although students said they understood the segmenting task, soon after commencing the 1st session uncertainties emerged about how to tell one design move from another and how to indicate their beginnings and their ends. Another uncertainty emerged as a consequence of the dialogue we employed in place of the monological think-aloud method for generating the protocol. Students were unsure about how to define a design move when the two designers talked of the same topic, and what to do when they interrupted each other. All difficulties that students reported serve us as significant clues to how future instructions need to be formulated. This might be our biggest challenge regarding the future of this research.
- We consider two training session were not enough for students to be familiar with the task they were asked to perform. Also the nature of these training sessions needs to make better use of training materials. Goldschmidt's examples served us well for a start. But we would need to introduce many more examples in a cycle of: asking observers to identify design moves; identifying their boundaries; getting immediate feedback of observed disagreements; and continuing with the segmentation of design moves until disagreements are tolerable.
- It is not enough to sample observers by criteria that seem adequate on logical grounds (design experiences, analytical skills, language competencies, stability of judgements, motivation to participate in the process, etc.). They also need to survive the training sessions by proving themselves to be reliable observers, interpreters, and judges of the stream of design moves they are confronted with.

Regarding the quality of data for analysis, we think the following would improve the generalizability of the conclusions:

- Reliability testing the improved instructions with control groups of trained observers from elsewhere: students from different universities; but also experienced linguists, content analysts, and design researchers. It would not be enough to leave advancing the analysis of design moves in protocols of design activities to students – graduates or post-graduates.

Ultimately, there has to be a validity check of the resulting segmentations and constructions of linkographies, ideally with the designers who generated the protocols of their design activity. We realize that this may be difficult as a linkograph is something potentially strange to a practicing designer. We would be satisfied if that designer would get additional insights from examining the linkographical results. We should question our analysis if that designer cannot

find him or herself represented in these results or see nothing of interest in them. We plan to further elaborate on and refine the rules for identifying design moves and assessing the reliability of linkographies. Our main goal is to help researchers obtaining reliable data for analyzing design processes. We are aware that human beings are not machines. When it comes to measurement, we should not expect the same precision a machine would achieve. But, on the other side, we should have some reliability associated to our data. The key contribution of further research in this area is the development not merely of appealing design concepts, but of concepts that can lead to replicable analyses that can be shared within the design community and lead to a better understanding and improvement of design activity. Our study took a first step in that direction.

Acknowledgements

The authors would like to thank Matthew Lombard, for his valuable insights and reference suggestions about calculating inter-coder reliability; Julio van der Linden, professor in the Departamento de Design e Expressão Gráfica at UFRGS, Universidade Federal do Rio Grande do Sul, who helped to tabulate students' data, the two anonymous reviewers, for their valuable comments, and most of all, the students who kindly volunteered to participate in the experiment.

References

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596.

Auld, F., Jr., & White, A. M. (1956). Rules for dividing interviews into sentences. *Journal of Psychology: Interdisciplinary and Applied, 42*, 273–281.

Austin, S. (2001). Mapping the conceptual design activity of interdisciplinary teams. *Design Studies, 22*(3), 211–232.

Ball, L. J., & Christensen, B. T. (2009). Analogical reasoning and mental simulation in design: two strategies linked to uncertainty resolution. *Design Studies, 30*(2), 169–186.

Ball, L. J., Onarheim, B., & Christensen, B. T. (2010). Design requirements, epistemic uncertainty and solution development strategies in software design. *Design Studies, 31*(6), 567–589.

Bilda, Z., & Demirkan, H. (2003). An insight on designers' sketching activities in traditional versus digital media. *Design Studies, 24*(1), 27–50.

Bilda, Z., Gero, J., & Purcell, T. (2006). To sketch or not to sketch? That is the question. *Design Studies, 27*(5), 587–613.

Cai, H., Do, E. Y.-L., & Zimring, C. M. (2010). Extended linkography and distance graph in design evaluation: an empirical study of the dual effects of inspiration sources in creative design. *Design Studies, 31*(2), 146–168.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics, 23*(1), 13–31.

Carmel-Gilfilen, C., & Portillo, M. (2012). Where what's in common mediates disciplinary diversity in design students: a shared pathway of intellectual development. *Design Studies, 33*(3), 237–261.
Chai, K.,H., & Xiao, X. (2011). Understanding design research: a bibliometric analysis of design studies (1996–2010). *Design Studies, 33*(1), 24–43.
Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: a practical guide. *Journal of Learning Sciences, 6*, 271–315.
Christiaans, H., & Almendra, R. A. (2010). Accessing decision-making in software design. *Design Studies, 30*, 641–662.
Craig, D. L. (2001). Stalking homo-faber: a comparison of research strategies for studying design behavior. In C. M. Eastman, W. M. McCracken, & W. C. McCracken (Eds.), *Design knowing and learning: Cognition in design education*. Elsevier.
Cross, N., & Cross, A. (1995). Observations of teamwork and social processes in design. *Design Studies, 16*(2), 143–170.
Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). The MIT Press.
Gero, J. S., & McNeill, T. (1998). An approach to the analysis of design protocols. *Design Studies, 19*(1), 21–61.
Gero, J. S., & Tang, H.-H. (2001). The differences between retrospective and concurrent protocols in revealing the process-oriented aspects of the design process. *Design Studies, 22*(3), 283–295.
Goldschmidt, G. (1991). The dialectics of sketching. *Creativity Research Journal, 4*(2), 123–143.
Goldschmidt, G. (1992). Criteria for design evaluation: a process-oriented paradigm. In Y. E. Kalay (Ed.), *Evaluating and predicting design* (pp. 67–79). Wiley.
Goldschmidt, G. (1995). The designer as a team of one. In N. Cross, H. Christiaans, & K. Dorst (Eds.). *Design Studies, 16*(2), 189–209.
Goldschmidt, G. (1997). Capturing indeterminism: representation in the design problem space. *Design Studies, 18*(4), 441–455.
Goldschmidt, G., & Weil, M. (1998). Contents and structure in design reasoning. *Design Issues, 14*(3), 85–100.
Guimarães, L. B. de M., Diniz, R. L., & Silva, S. A. (2002). Design participativo: o caso do posto de vendas em loja de departamentos. *Anais P&D Design*.
Kan, J. W. T., & Gero, J.,S. (2008). Acquiring information from linkography in protocol studies of designing. *Design Studies, 29*(4), 315–337.
Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Sociological Methodology, 25*, 47–76.
Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and Quantity, 38*(6), 787–800.
Krippendorff, K. (2011). Agreement and Information in the reliability of coding. *Communication Methods and Measures, 5*(2), 93.
Krippendorff, K. (2012). The reliability of distinctions. Personal communication 2012.7.31.kkrippendorff@asc.upenn.edu.
Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Thousand Oaks, CA: Sage.
McNeill, T., Gero, J. S., & Warren, J. (1998). Understanding conceptual electronic design using protocol analysis. *Research in Engineering Design, 10*, 129–140.
Tang, A., Aleti, A., Burge, J., & van Vliet, H. (2010). What makes software design effective? *Design Studies, 31*(6), 614–640.

Valkenburg, R., & Dorst, K. (1998). The reflective practice of design teams. *Design Studies, 19*(3), 249−271.

Van der Lugt, R. (2001). *Sketching in design idea generation meetings*. Doctoral dissertation, Delft University of Technology.

Yalçinkaya, S. Í. (2010). *An inter-annotator agreement measurement methodology for the Turkish discourse bank (TDB)*. Dissertation, The Graduate School of Informatics, METU − Middle East Technical University.