Understanding the meaning potential of pictures in speech and language assessment

CATHERINE E. BROUWER, CHRISTIAN M. JOHANNESSEN & MARIT C. CLAUSEN

University of Southern Denmark, Department of Language and Communication

ABSTRACT
The goal of this article is threefold: (1) To question the use of images as straightforward prompts in language tests in relation to test validity (2) to show how the available documentation for tests in current use in Denmark reveals that test designers generally overlook the design of pictures in test development, and (3) to present an exemplary analysis, generally based on Kress and van Leeuwen's (2006) metafunctional approach, demonstrating the meaning potential of pictures used in tests. The tests we consider are materials for measuring children's linguistic abilities. Firstly, we demonstrate that pictures in language assessment procedures is under-researched and rarely considered in the validity evidence for testing materials. Secondly, we demonstrate the usefulness of a metafunctional approach when assessing the aptness of a picture for a given test situation.

KEYWORDS
language assessment; social semiotic analysis; pictures; validity evidence; test development

1. Introduction

It is beyond question that pictures are ubiquitous in practices of language education practice and language testing. Pictures are convenient and cheap to produce, reproduce and distribute. In testing of linguistic ability, picture naming is seen as a useful and in some cases the best procedure (see e.g. Haman, Luniewska & Pomiechowska 2015). Pictures seem difficult to avoid for the foreseeable future, and so it stands to reason that we should continue to develop the best possible practices for using pictures in language tests. This article focuses on pictures used in material for measuring children's linguistic abilities. The article has three separate but related goals. First, we question the use of images as straightforward in the context of measuring children's linguistic abilities on the basis of different strands of earlier research. Second, we scrutinize a number of tests that employ pictures to look at validity evidence in relation to this particular aspect of the tests. Finally, based on our argument in the first sections of the study, we point at a framework for analysis of test pictures that may contribute to providing validity evidence for the quality of these pictures in a systematic way, illustrated with a discussion of pictures used in the *Expressive One-Word Picture Vocabulary Test (EOWPVT)* (Martin & Brownell 2013).

rineke@sdu.dk

In pedagogical and educational practice, a distinction is made between screening and test materials, that both belong to the category of assessment, which also includes less formal procedures. Screening materials are used to monitor children's general language development with the purpose of early language intervention. Furthermore, more precise diagnostic procedures in speech and language therapy usually rely on tests of the child's communicative and linguistic ability (in concert with anamnesis, observations of spontaneous speech, and assessment of general development) (Nettelbladt, Samuelsson, Sahlén & Ors 2008), making assessment crucial. A comprehensive range of material is available for clinicians. Most of it has been designed by experts in psycholinguistics, psychology, linguistics, and speech and language therapy, with the aim of diagnosing a range of language disorders. In a Danish context, materials that have been developed in Denmark, such as *Kaj-testen* (Dahms, Bergstrand & Nakskov 2005) and *Viborg Materialet* (Pedersen & Kjøge 2000) as well as adaptations or translations of material developed in English-speaking countries, such as the *EOWPVT*, are in widespread use in both general and specialized practice.

No matter if a material is developed for screening, test, or assessment purposes, whether internationally published or developed and used locally, and regardless of scope, depth and precision, more often than not such materials rely on pictures as a vehicle in the assessment procedure. While the objective of the test procedure is to give insight into the child's linguistic development, the procedure requires the child not only to perform linguistically, but also to recognize pictured objects, events or scenes. Recognizing the picture is every bit as much a task as the construct under consideration, since it equally requires processing and relies on earlier acquired knowledge (Ward 2006). Since its function in the test is rather a vehicle of operationalization, however, the question remains whether picture recognition can be seen as straightforward in children.

2. Children's understanding of pictures in a linguistic test situation

Few studies exist that focus on children's recognition of pictures per se. Different strands of research consider linguistic expression in relation to pictures. Studies within e.g. the paradigm of 'Thinking for speaking' (Slobin 1996) employ picture books as elicitation for children's linguistic expression. Although there are probably important indications about picture understanding to be found in the data of these studies, the focus of this strand of work is on the production of spoken language, rather than on the understanding of the pictures, and in the studies themselves, the pictures are typically regarded as a straightforward point of departure.

Studies on visual literacy are concerned with how information is presented and understood in the form of images. Within this field, most studies focus on the understanding of narratives through pictures, the understanding of text with the aid of pictures, or art appreciation, typically for school children or older subjects. However, a small number of studies on how young children understand images exists. Walsh (2003) e.g., in a qualitative study that considers young children's responses to reading a picture book, categorizes some of the responses as labeling

(naming the item that is depicted) whereas others can be seen as a range of different types of comments, which make sense in the context of shared reading. However, this study, as well as other studies on children's visual literacy, was done with regard to responses to pictures in a picture book, and not with the objective of studying how children name or label pictures.

Based on Vinberg Hansen (2017), who describes how children may interpret the pictures in two different tests, Table 1 illustrates alternative labels made by children taking the Danish test *Sproglydstest* (Kaufmann & Ege 1974)[1].

TABLE 1: ALTERNATIVE INTERPRETATIONS OF PICTURES IN *SPROGLYDSTEST* (KAUFMANN & EGE 1974)	
Target word	**Produced word**
risengrød ('rice porridge')	*nisse* ('elf')
strømpe ('stocking')	*sok* ('sock')
kødsovs ('bolognese sauce')	*ketchup* ('ketchup')
pasta ('pasta')	*mad* ('food')
smør(e) ('spread')	*honning* ('honey')

The word *strømpe* ('stocking') is close to being synonym with *sok* ('sock')*, mad* ('food') is a super category for the target word *pasta* ('pasta')*, honning* ('honey') is replacing an object for the target of the verb *smøre* ('spread')*,* and *ketchup* ('ketchup') may be the child's preferred sauce eaten on pasta rather than *kødsovs* ('bolognese sauce'). It is interesting that these alternatives seem reasonable because, one way or another, they are all related to the target words. In these cases, thus, it appears that the failure to offer the targeted word does not result from ambiguities in the design of the pictures themselves, but from a host of other factors. Among them are, to name a few, not knowing the target word, the child's range and depth of linguistic knowledge, being distracted, the child's understanding of the objective of testing, or (semantic) priming effects. Better insight into the child's linguistic ability would be possible if one could exclude the quality of the pictures as a factor for failure to produce the target word.

2.1 Studies focusing on picture recognition

We have been unable to identify studies, apart from Vinberg Hansen (2017), that deal with children's recognition of pictures in the context of linguistic assessment.

Broadening the scope to different fields, we find studies that treat picture recognition within cognitive psychological assessment. Most of these studies are based on Snodgrass & Vanderwarts 1980 study, in which they presented a set of pictures to be used in assessment, and which were standardized in relation to four variables: Name agreement, image agreement, familiarity

---

1   Vinberg Hansen chose an older version of this test since this was the version used in the specific practice which was studied.

and visual complexity. Subjects were asked to name the picture shown, resulting in scores for name agreement between subjects, and they were asked to rate the extent to which the shown pictures exhibited recognizable objects, whether these objects were usual or unusual, and how complex they judged the picture to be. The findings were that the intercorrelations of these variables were quite low. Furthermore, in accordance with the examples of the Vinberg Hansen study above, this study found that when pictures were not named with a word that had a high name agreement, the subjects would use super- or subcategories, add a modifier, use a synonym or a coordinate (e.g. *avocado* for *artichoke*). The subjects however also had the possibility to indicate that they did not recognize the concept of the picture, that they did not know the name of the item indicated or that they experienced a tip-of-the-tongue phenomenon. A number of studies have replicated the Snodgrass & Vanderwarts study, either based on the same set of pictures or a newly developed one (e.g. Brodeur, Dionne-Dostie, Montreuil & Lepage 2010), investigating the same variables or additional ones, while a larger number of studies investigate the variables in relation to a population with other mother tongues than English (e.g. for Icelandic (Pind, Jónsdóttir, Gissurardóttir & Jónsson 2000)). As a whole, this strand of research establishes a norm of what the most common label is for each particular picture amongst a population, thus making available a set of standardized pictures for research and assessment.

The strand of research described in the above employs adults as its subjects. As Cannard, Blaye, Scheuner & Bonthoux (2005) as one of the few studies on children's responses to pictures have shown, children respond to pictures in different ways than adults. The study operated with the same set of pictures as Snodgrass & Vanderwart (1980), and the subjects were French children between 3 and 8 years, grouped according to age. The study showed significantly less name agreement between child subjects than between adult subjects, and it was lowest for the youngest group of subjects. Young children tend to respond with 'I don't know' much more often than older children, and children as a whole again much more often than adults, indicating that they either do not know the concept or do not recognize the picture. Furthermore, unexpected naming[2] was much higher for children than for adults (with the group of youngest children the highest). While some of the answers were categorizable as super- or subcategories, synonyms, coordinates or with the addition of a modifier, children in this study responded often with names that were judged to be wrong. Finally, the study indicated that children prefer giving a wrong answer over admitting that they do not recognize the object. The study concludes that "[i]n order to select pictoral stimuli for studies with children, it is important to choose pictures that are correctly identified by children" (Cannard et al. 2005: 423).

2.2 Implications of picture recognition studies to linguistic test development
What may be concluded from these different types of studies, is that it cannot be assumed that pictures can be used as unproblematic means for investigating or assessing children's linguistic behavior. Naming a picture requires that the depicted item is recognized, and, according to

---

2   In the studies, responses that were not high in name agreement, but also were not super- or subcategories, modified, synonymous, or coordinate responses.

most neuroscientific models, this takes place at an earlier stage than the actual naming (even though contemporary models allow for interaction between stages of recognition and naming) (Ward 2006). The trouble is that in the methodologies in the paradigms mentioned above, the recognition of the pictures is intertwined with the ability to name what it depicts, creating a chicken or egg causality dilemma. Either the studies focus on recognition of pictures and employ linguistic means in order to document whether and how such recognition has taken place, or the studies focus on linguistic abilities and assume pictures as a more or less straightforward means to be able to describe it.

It is however not the case that test-developers are not aware that the quality of pictures may pose problems for test-takers. The majority of testing materials prescribe ways for test administrators to deal with non-target versions of picture-naming. One might argue, therefore, that it is unproblematic when test prompts are misinterpreted. There are, however, three reasons why this is not so.

First, it is favorable when a child spontaneously gives the response intended in the test design. Using various prompts is less favorable, because it might give a less clear picture of the child's actual ability (Goldstein, Fabiano & Iglesias 2004).

Second, misinterpretations may in some cases lead to reformulations, which make the test take longer. Not only is this more expensive, it can also cause the child to be bored and more easily distracted.

Third, although unsupported by evidence, one can worry that clinicians may sometimes score an answer as correct, since it makes sense, as in the produced words in table 1, even when the word produced was not the targeted one. We have been told that this sometimes happens even when a standardized test is employed, which would seriously jeopardize the validity of the result (see section 3). For these reasons, the quality of the design of pictures for tests is an issue.

Features of a picture's design may be the only reason why a child fails to produce a targeted word. Put differently, we cannot *a priori* assume that a child will recognize any depiction of a concept (word) even if the child knows the targeted word. Visual configuration of pictures may be a confounder for test results, and it concerns the validity of the test.

In section 3 we will demonstrate how developers deal with the pictorial aspect of test design, including how they document it to test administrators and users. This part of our argument concerns the extent to which validity evidence regarding depictions in test and assessment materials is provided. In section 3 we depart from Kress and van Leeuwen's multimodal social semiotics (2006) to propose a way of thinking about picture design that covers all significant aspects of their meaning potential as an approach to validating this aspect of tests.

3. Validity in test materials

Validity in relation to testing concerns the degree of probability that the test measures what it is supposed to measure, e.g. children's vocabulary or phonological abilities. As outlined above, we question whether tests that employ pictures necessarily measure such linguistic abilities as opposed to testing the abilities to recognize pictures.

Validity is typically not understood as an inherent quality of a test, but rather something that can be demonstrated by the production of evidence (AERA, APA & NCME 2014)[3]. In many test situations, children need to respond linguistically to a task which relies on successful understanding of a picture. When children name, sort or point out pictures in tasks that are developed to measure their linguistic abilities, can they be assumed to readily understand those pictures? Or would it be possible that their naming, sorting or pointing out is directly reflecting their understanding of pictures – rather than their linguistic abilities? The understandability of pictures in test materials thus concerns validity, a parameter that test developers may concern themselves with when designing a test.

Designers' expertise with test development varies greatly in tests in common use in Danish speech and language therapy, ranging from those developed by researchers, by professionals with special competencies in test development, to those developed by practitioners of the field. However, test development is also a field of expertise in its own right, which is intimately tied to the study of psychometrics (Black 1998; Raykov & Marcoulides 2011). In some areas, such as education, professionals can make a career in test development. Within the field of test development, the quality of a test is in part determined by its adherence to overall standards in test development as described by e.g. American Education Research Association (AERA et al. 2014). These standards concern the quality of the test and include descriptions of how reliability and validity evidence of various kinds can be produced. In this field, it is common to document the development process, including the production of validity and reliability evidence, either in research publications (e.g. Bishop 1977; Lavesson, Lövdén & Hansson 2018; Letts, Edwards, Schaefer & Sinka 2013) or in test manuals.

Thus, some of the tests used in speech and language therapy have been designed in accordance with test development procedures to heighten their reliability and validity. These procedures are sometimes described in guidelines, best practice descriptions, in the aforementioned standards, or in psychometric literature. The development of a number of tests in the field only follows procedures in part, or, seemingly, not at all, as we will develop below.

One such set of standards is presented by the National Council on Measurement in Education, an organization in the US that specifically provides standards to professionals, including developers, administrators, scorers, interpreters and educators on assessments: *Code of Professional Responsibilities in Educational Measurement*. Several professional organizations in the US cooperated on developing what is called *The Code of Fair Testing Practices in Education* – this code is sponsored by American Speech-Language-Hearing Association (ASHA). The code describes, amongst others, standards for test developers and administrators. A more elaborate version of this code is the *Standards for Educational and Psychological Testing* from the American Educational Research Association (AERA et al. 2014). An international, more specific standard for language testing, are the *International Language Testing Association (ILTA) guidelines*. All of these guidelines include a section requiring test developers to follow procedures for securing reliability and validity.

---

3   American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Concluding from the above, there are two principal ways in which the quality of test items, and more specifically the quality of pictures that are included in a test can be evaluated: By considering procedures of test construction, and by considering any measures of the validity of the pictorial aspect of a test obtained either during test construction or afterwards. These are of course mutually dependent, but can be analytically distinguished. We will discuss test construction in section 3.1 and validity measurement in section 3.2.

3.1 Test construction: Four different approaches

Ryan, Lopez and Sumerall (2001) describe four distinct approaches to test construction:

- a. *The criterion-keyed approach*, which starts out with a definition of the construct (e.g. vocabulary size, phonological skills, morphological skills, etc.). A sample of items is then administered to different categories of people, e.g. age groups, in order to find out whether and in what way responses can be scaled in relation to these categories.
- b. *The analytic approach* starts out with a number of items, which are based on a theory, e.g. a theory of phonological development, and administered to a large sample. The results are treated in factor analysis (a statistical method) and interpreted in relation to the theory that guided the development of the items. This approach needs a strong theoretical basis. Replication is thought to be necessary in order to secure stable results.
- c. *The rational/deductive method* is a combination of the two first approaches. Initially, items are selected that are thought to represent a specific domain, based on theory. Observations and expert evaluation may be used to secure adequate sampling of the domain. The items are then administered, typically, to a normative sample of a population and results are analyzed and interpreted, again in relation to the theory. A large pool of possible items is then reviewed by experts in order to secure cultural appropriateness, understandability and content. Finally, one or several pilot studies are conducted on a normative sample of individuals, which may give rise to the exclusion of items.
- d. *An item analysis* may include a review of test content and specifically addresses item difficulty, item discrimination, and item response theory. The latter three concepts are analyzed based on the administration of items to a sample and calculating different correlations, e.g. those between items responses, or those between item response and composite score.

Based on this description, we may conclude that, only in the rational/deductive approach to test construction, a central aspect of the construction concerns expert evaluation of items. In all approaches, the focus is on quantitative analysis of item responses in a tentative version of the test-under-construction.

3.2 Validity measures

What comprises validity of a test has changed somewhat over the years. Previously test developers used a variety of validity types, such as construct validity or content validity etc. Today, validity in a testing context is seen as a comprehensive and unitary concept that is informative of the appropriateness and usefulness of a test in relation to a specific purpose (AERA et al. 2014). It includes not only test characteristics but also appropriateness of use and the usefulness of the interpretations of test scores for the designated context.

In order to judge the validity of a test, several types of evidence are thought of as essential: evidence based on test content; evidence based on internal structure; evidence based on relations with other variables such as test takers' age and intelligence scores; evidence based on (test taker) response processes; and evidence based upon the consequences of testing (Geisinger 2016).

These five types of evidence are thought of as essential to be included in test specifications or test manuals. More specifically, the aforementioned rational-deductive approach to test validity includes a reasoning statement of the proposed interpretations and uses of a test (a claim: <u>what</u> is the objective of this test?) as well as an evaluation of this statement/claim (<u>how</u> can this test fulfill this objective?) (Kane 2016). The statement would include a description of inferences that can be made on the basis of the test scores as well as a specification of assumptions about, for example, what language is, what intelligence is etc. The evaluation would provide supporting evidence for the inferences as well as for the assumptions. Importantly, we observe that also with respect to validity evidence a predominance of quantitative methods can be observed, just as was the case with test development approaches.

The present study concerns evidence based on test content, which refers to the themes, wording, and format of items, tasks and/or questions included in a test, and focuses on the relationship of these and the construct the test is intended to measure (AERA et al. 2014). According to Ryan et al. (2001), content validity evidence is obtained by qualified reviewers who carefully produce a formal evaluation. This process, then, is one of the few in which a more qualitative analysis is conducted. Observe, also, that from the latter description of what comprises validity evidence, the concept of face validity is not included. The description of what covers content validity seems to emphasize that items should be considered in a systematic way, and by experts.

From the above treatment of test quality in terms of development approach and validity measures, it follows that clear definitions, and international guidelines are available – and these could also be employed in relation to the pictorial aspect of tests.

In the following, we report on a number of assessment materials, concerning children's communicative and linguistic abilities, particularly whether these materials' manual includes information on the approach of test development, and whether the applied approach covers expert evaluation of specifically the pictures used in the test. Furthermore, we report on the inclusion of validity evidence based on test content, with regard to validity of pictorial aspects of the test items.

3.3 Evaluation of test materials with regard to validity evidence of pictures

We scrutinized the information available for a number of tests that utilize pictorial information and are considered to be widely used today in Denmark for the evaluation of children's linguistic abilities, (see table 2). Some of the information is available from the test manual or test specifications, and sometimes scientific articles provide such information, and these may actually concern the original test materials rather than a translated or adapted version. Neither information regarding systematic approaches to test development as the ones described in Geisinger (2016; see section 3.2.), nor information that makes it easy to conclude which of these approaches is applied, is made available for test users in any of the tests considered. Regarding validity evidence for pictorial aspects, based on our inspection of background literature on the specific tests (either in a manual or in research publications) we identify four categories of explication with regard to picture materials: 1. no information; 2. asserting face validity; 3. decisions to exclude items are made on the basis of statistical analyses of a trial (item analysis); and 4. item analysis based on a trial (as in category 3), as well as some systematic planning of pictorial aspects of the items. The results are represented in Table 2.

TABLE 2: CATEGORIES OF EXPLICATION REGARDING TEST MATERIALS USED IN DENMARK	
Categories of explication	**Materials**
1. No information	*Kaj-testen* *Sproglydstesten Metha* *Dansk Impressiv Morfologisk test (DIM)* *Sproglig Test 1, 2, 3, 4*
2. Asserting face validity	*Viborg Materialet* *Expressive One-Word Picture Vocabulary Test (EOWPVT)* *Test of Reception of Grammar (TROG-2-Danish adaptation)*
3. Decisions to exclude items made on the basis of statistical analyses of the trial (item analysis)	*New Reynell Developmental Language Scales (NRDLS - Danish translation)* *Logopædisk udredning af Fonologiske Vanskeligheder (LogoFoVa)*
4. Employs trialling and item analysis (as category 3), as well as some systematic planning on picture production	*Peabody Picture Vocabulary Test, fourth edition (PPVT-4)* *Clinical Evaluation of Language Fundamentals, fourth edition (CELF-4 - Danish adaptation)*

As can be seen, some test materials provide no information on how pictures were dealt with (category 1) or the test developers asserted face validity, typically by asserting that pictures were chosen that were thought to be recognizable by children (category 2). The 3rd and 4th category of test include a trial: A first version of the test was developed and tried out on a smaller sample of children – the results of these trials were then statistically analyzed which pointed at prob-

lematic items. Subsequently, such problematic items were typically excluded from the material. This was the case for e.g. the *New Reynell Developmental Language Scales (NRDLS)*. An extensive report of test validation was provided in Letts et al. (2013). Less extensive articles are available on earlier versions of this test (*RDLS III*) (Edwards, Garman, Hughes, Letts & Sinka 1999). In the latter, the authors assert:

> The pictures and toys used in the test were carefully trialed to ensure that they were attractive and familiar to children; materials that proved difficult for children to recognize at the trialing stage were removed from the final version of the test. (Edwards et al. 1999: 161)

Such trialing essentially entails a quantitative measure. In this case, item analysis was employed: This is basically providing a measure of individual items compared to a total test score. By statistical analysis, it is possible to find out if some individual items stick out in terms of difficulty in relation to others. In the case of *NRDLS*, thus, individual items were evaluated as in terms of difficulty in relation to the overall test, i.e. the other items. However, the question then is what 'an item' consists of. Is it the picture in itself, or, alternatively, the targeted linguistic expression? Is it the picture in *combination* with the targeted linguistic expression? Is it the *task* of selecting a picture upon hearing a linguistic expression – including whatever other information there is available in the test situation such as gaze, intonation, gesture etc. from the test administrator? In essence, of course, a trialing includes all of these. In our paper here, however, we are interested in focusing on only the pictorial aspect of the test material, as we suggest that these in themselves may exhibit different degrees of recognizability

As an example of category 4, *PPVT-4* (Dunn & Dunn 2007) asserts in its manual that pictures were chosen to be recognizable for persons that may have moderate visual disabilities and carefully selected to represent the current US population (race/ethnicity, sex, culture). The original items including pictures were field tested and revised during several rounds. Subsequent editions included more items. They were typically identified as suitable by way of item analysis and Rasch item analysis. The 3rd edition of *PPVT* (Dunn & Dunn 1997) included a rational/deductive approach to item analysis by use of a panel of experts scrutinizing for pictures being offensive or biased with regard to sex and/or race/ethnicity, which informed also the later *PPVT-4*. For the *PPVT-4* item pool, a number of design requirements were set up for the pictures, mainly addressing issues of use of color and contrast. A graphic designer edited the pictures. Also, for *PPVT-4*, quantitative pilots and try-outs were employed to scale items according to difficulty and rule out problematic items. Reviewers typically had a background in speech and language therapy. The available evidence concerning validity and reliability for this test, including the pictures, is substantial. Included are (a) the initial evaluation by the test developers of items as 'easily illustrated', (b) the graphic designer's editing work conforms to a design manual addressing issues of color and contrast, (c) a bias panel's scrutinizing for offensive or biased items, and (d) elimination of items based on quantitative evidence, mainly item analysis.

Concluding, evidence for the interpretability of pictures is not provided for some of the ma-

terials we have investigated. In other cases, face validity is asserted, or validity is based on a quantitative argument. Only in a few cases, a strategy regarding the selection of pictures is employed.

In no case, analysis has been conducted that systematically scrutinizes the quality, and thus suitability of pictures and illustrative strategies, indicating whether and how pictures are understandable in terms of their configuration.

4. A metafunctional approach to pictures in language tests

The practices concerning validity of pictures in test development as described above is problematic in more ways than one. The largely quantitative criteria for choosing pictures for concepts exhibited by, for example, *NRDLS* (Edwards et al. 2011) and *LogoFoVa* (Clausen 2014), may indeed indicate that a picture is problematic for some reason, and so should be omitted from a test, but not why it is a problem. For this reason, the professional community is cut off from an opportunity to learn, to collectively be smarter about using pictures in tests and better anticipate which of their aspects will be problematic and which will not. A way to get around this is by developing (1) a qualitative understanding of problematic pictures and (2) a meta-language with which we can address problems. This section of the article suggests, therefore, a way of inquiring systematically into any aspect of meaning potential that may impinge on a picture's aptness for a test.

4.1 The social semiotic notion of metafunction

As a broad, guiding heuristic, we use the term *metafunction*, which is one of the theoretical tenets of Kress and van Leeuwen's social semiotic work (2006: 40-42) on visual communication, in a first attempt to cover the relevant aspects of a picture's meaning potential. This again continues the legacy of late British linguist Michael Halliday and his Systemic Functional approach to language (Halliday & Matthiessen 2013). One of Halliday's many contributions to contemporary descriptions of communication is the idea that people cooperate to simultaneously make three different kinds of meaning, all of which are necessary for communication to function properly, when they interact. They do so regardless of whether their interaction is face-to-face in the here and now or discontinuous over time and distance and mediated by texts. He called these three potentials for meaning the ideational, interpersonal and textual metafunctions.

For example, a child is shown a picture of a dentist and patient in a dental clinic (like Plate 63 in the *EOWPVT* (Martin & Brownell 2013)) in a test of the child's ability to name things, concepts, and actions. The picture will inevitably be meaningful for the child in a host of ways that may or may not be relevant for the construct of the test. She may see the picture as a picture of a dentist, which is the test item's target word, and provide a correct answer. Or she may see the picture as a picture of a patient, or indeed of a dental lamp, and provide an incorrect answer.[4]

---

4   Of course, she will probably make some meanings rather than others. She is unlikely to respond with *bath* or *breakfast*. And test designers do have resources at hand to guide her meaning-making along some trajectories rather than others. But it is beyond dispute that Plate 63 also depicts a patient and a lamp, even if these are

However, regardless of what she sees the picture as a picture of, it will also be meaningful in other ways. By simply having a specific style it may engage her, leave her indifferent, or repel her. It may have a fun, silly, serious, or boring look that she is used to from books she reads or that are read aloud to her, from things that parents or teachers find important, from things that, in her experience, are meant to entertain, educate or warn her. And, regardless of what is depicted, and how its look makes her feel, she will pick up meaningful patterns in the picture – or between this plate and the others in the test – of how, for example, lines, color, shape, size, space, and frames consistently suggest that traits of it are important, different, alike, belong together, are separate etc. As will be demonstrated, each of these broad areas, within which the child makes meaning with the picture, can be more or less conducive for the purpose of the test and, in some cases, may be suspect as confounders.
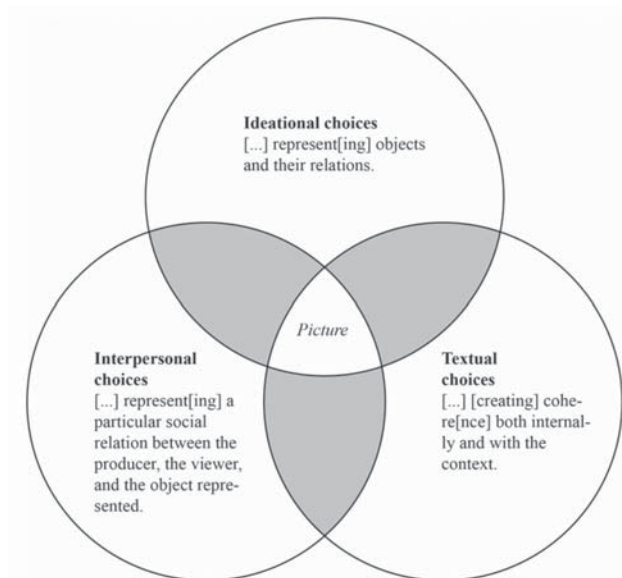
The example serves to give a first impression of how social semiotics distinguishes the ideational, interpersonal, and textual metafunctions. When the child sees the picture as a visual construal of objects in her experience, and their relation, she is making ideational meaning. When she enacts social relationships through the picture (or maybe rather when the designer enacts social relationships upon her through the picture), she is making interpersonal meaning. And when she experiences the picture as a cohesive unit, a text, either in its own right or as part of a set, she is making textual meaning.

The meaning potential of any picture for a specific test taker is of course at the heart of the matter and can only be analyzed in the specific context. However, test developers as well as test consumers may be made aware of the choices for guiding and constraining test takers' meaning making process along trajectories appropriate for the test purpose. A key idea in social semiotics is that semiotic resources (including speech, writing, image, color, typography, to name but a few) evolve and are shaped over time through the social work of members of a given culture, and they are described as interrelated systems of paradigmatic[5] choice relations. Crucially, some choice relations are meaningful in an ideational sense whereas others are meaningful in interpersonal or textual senses. It follows, that the dentist picture can be analytically regarded as the sum total of choices made by the designer within an array of choices offered her by ideational, interpersonal, and textual systems, and that for each choice made there are a host of potential choices not made.

---

considered context objects, and it is not entirely impossible that the child's meaning-making latches onto those rather than the dentist.

5   Paradigmatic choice should be understood here in the sense of patterns of "what could go instead of what" as opposed to syntagmatic choices of "what can go with what" (Boeriis 2009: 217-224).

Figure 1: A given picture is an intersection of ideational, inter-personal, and textual choices made by the designer within arrays of choices (derived from Kress & van Leeuwen 2006).

Ideational choices
[...] represent[ing] objects and their relations.

Picture

Interpersonal choices
[...] represent[ing] a particular social relation between the producer, the viewer, and the object repre-sented.

Textual choices
[...] [creating] cohe-re[nce] both internal-ly and with the context.

4.2 Conceptual Specificity

Where the validity of a test using pictures to assess a linguistic construct is concerned, test developers must take precautions to ensure that a failure to produce the target response is an effect of the child's linguistic ability, not the fallibility of its interpretation of the picture. American visual semiotician Steven Skaggs' Peirce-inspired notion of conceptual specificity (2017: 91) holds that any picture offers a range of possible interpretations on a scale of potentiality between absolutely certain denotation or connotation (an "ideal and unreachable endpoint" (Skaggs 2017: 93)) to absolutely uncertain (an equally ideal endpoint). Simply put, the design of the picture can act to narrow or broaden the scope of its meaning potential. In the following three subsections, we will employ simple commutations to demonstrate possible consequences for ideational, interpersonal, and textual specificity of different choices made in a depiction of the concept 'dentist'.[6]

4.2.1 IDEATIONAL SPECIFICITY

Ideational choices offer "different ways in which objects and their relations to other objects can be represented" (Kress & van Leeuwen 2006: 42). A profession like 'dentist' is fairly abstract. In depicting it, we tend to rely on someone performing the actions of the trade, usually with typifying tools in or at hand. Figures 2-4 depict a male dentist working on a female patient with a curette in his right hand and mirror in his left, not unlike Plate 63 in the *EOWPVT*.
Figure 2 is low in specificity of the concept but, given the right context, it could pass as a picture

---

6   In order to avoid infringing on others' intellectual property, none of the pictures in this article can be found in actual language tests. The authors have produced all of them for illustration purposes.

Figure 2: A picture, which ambiguously depicts a dentist and patient

Figure 3: A picture that offers more cues about cultural types and narrows the scope of possible interpretations

Figure 4: An ideationally highly specified picture of a dentist and patient

of a dentist. It probably requires the right interpretation of what goes on around the patient's mouth, and the material affordances of the drawing are not very helpful: It could show anything from what was intended to e.g. something involving first aid to someone having swallowed an object.

Figure 3 is less ambiguous because a choice has been made to depict the dentist with typifying, so-called, "possessive attributes" (Kress & van Leeuwen 2006: 87): A shirt, lab coat, and surgical gloves on the dentist and a bib on the patient all cue us in on dentistry as a social practice, even if what goes on still looks like it could involve shoelaces. Figure 4, adding further cues to make any other interpretation unlikely, is a highly ideationally specific picture of a dentist.

4.2.2 INTERPERSONAL SPECIFICITY
Interpersonal choices enable us to specify "a particular social relation between the producer [of a sign], the viewer and the object represented" (Kress & van Leeuwen 2006: 42). Figures 5-7 demonstrate three vastly different styles of depicting, more or less, the exact same ideational structure and composition.

Figure 5: A formal, technical style of illustration, found in e.g. user manuals or airline safety cards

Figure 6: A sensory and atmospheric style of illustration, which is found in e.g. graphic novels

Figure 7: An informal, humoristic style of illustration, which one finds in e.g. school books and comics

It seems clear that they target different audiences and also that they express different identities on the part of the producer. Figure 5 looks like a picture in a user manual. It specifies a mode of address accepted among adults in situations characterized by sobriety, objectivity and authority. Figure 6 looks more like an illustration from a graphic novel. Its shading specifies a sensorially evocative and atmospheric mode of address, which would maybe harm its credibility in many professional contexts, but which is appropriate in e.g. fiction. Figure 7 employs dynamic strokes, flat shading, and draws on pictorial conventions from comics (e.g. facial features). Its mode of address is typical in informal situations characterized by humor, like in newspaper comic strips. It is a common style when adults address children, probably under the assumption that it will engage them.[7]

4.2.3 TEXTUAL SPECIFICITY

Finally, textual choices enable us to create texts, "[...] complexes of signs which cohere both internally with each other and externally with the context for which they were produced" (Kress & van Leeuwen 2006: 43).

---

7   Further presentation of interpersonal analysis, regarding "modality" can be found in Boeriis (2009: 269-298) and Kress & van Leeuwen (2006: 154-163) and regarding "coding orientation" in Kress & van Leeuwen (2006: 163-174). Also, the meaning potentials of color (Kress & van Leeuwen 2002) and graphic strokes (Stötzner 2003) are relevant topics here.

Figure 8: In this scene from a dental clinic, stroke weight and grey scale make the dentist most salient

Figure 9: Here, the patient is most salient

Figure 10: And in this picture, the lamp is more salient than both patient and dentist

Figures 8-10 demonstrate how a structural contrast (in this case between two line weights and two shades of line color) can be used to make three pictures, which are identical in every ideational and interpersonal respect, mean different things through different so-called "salience" (Kress & van Leeuwen 2006: 201). Figure 10 is less a picture of a dentist or patient than a picture of a dental lamp, because the picture's salience hierarchy construes the lamp as worthier of our attention than the dentist or patient (who in this case seem to be relegated to acting as contextual props for heightening the ideational specificity of the dental lamp).

This is just a small taste of what understanding the textual metafunction can offer to developers of speech and language tests. Again, we refer generally to Kress and van Leeuwen (2006: 201-203) and Boeriis (2009: 210-216) for a fuller picture of "salience". Also, the textual meaning potential of "rank" and "affinity mechanisms" (Boeriis 2012) and, to a lesser degree, "information structure" (Boeriis 2009; Kress & van Leeuwen 2006: 179-201) is relevant for understanding how pictures in tests actually succeed (or fail) in specifying the precise meaning potential required by a test construct.

4.3 A discussion of visual improvement in the *EOWPVT*

The above brief introduction to a metafunctional approach to pictures only does justice to the most general insights from social semiotics and Halliday's and Kress and van Leeuwen's contributions. This section discusses the strategies used in the *EOWPVT* (Martin & Brownell 2013) to overcome common challenges with ideational specificity of pictures. More specifically, we observe a kind of visual improvement in many of the more conceptually and structurally more complex pictures.

Figure 11: Reconstruction of Plate 63, which uses context to further specify an abstract profession, 'dentist'

Figure 12: Reconstruction of Plate B, which refers to 'toe' in the context of 'foot'

Figure 13: Reconstruction of Plate 25, which uses context to further specify an amorphous object, 'pillow'

Generally speaking, the test's developers seem to have found it much easier to pictorially specify some ideational meanings than others. This appears to be the case when pictures specify concrete objects of our everyday experience, when they specify meronymic wholes rather than parts, and when they specify objects with a distinctive, easily recognizable shape. Among examples in EOWPVT are Plate A ('dog'), Plate 1 ('apple'), and Plate 12 ('banana').

In other cases, however, the developers seem compelled to provide (sometimes quite a lot) of visual context in order to heighten the ideational specificity of a picture. This appears to be the case in depictions of (i) concepts that are not objects of our concrete experience (abstract) as in Figure 11 ('dentist'), (ii) parts of wholes (meronymy) as Figure 12 ('toe'), or (iii) when objects are indistinctly shaped (amorphous) as in Figure 13 ('pillow').

This strategy, reasonable though it may be, entails losses as well as gains. Although developers may gain ideational specificity by depicting, for example, the indistinctly shaped pillow with someone resting their head on it, they lose textual specificity understood as cohesion with the context of the test situation and its requirements for clarity. The child may latch onto the wrong part of the picture and fail to produce the word *pillow* – not owing to its own linguistic inability but because of the way the picture was designed.

In such cases we observe four different strategies for improving the ideational clarity of the picture. These all serve to make the target for that test item more salient in the overall composition of the picture. We sometimes observe more than one of these used in a single picture. They are:

1. Contrast in stroke (line) weight as a salience device
2. Contrast in color saturation as a salience device
3. Framing as a salience device
4. Indexing with arrows as a salience device

Examples abound in the *EOWPVT*, for example, Figure 14 illustrates Plate 28 with the target word *hair*. Like 'toe', 'hair' is a meronymical part of the body and it is also fairly indistinctly shaped. Note how developers have used both stroke weight and color saturation as well as an arrow to specify that the target word is *hair* and not *face* or *head*. Figure 15 illustrates Plate 177 with the target word *isthmus*. Here, in addition to using an arrow, developers have framed the isthmus as if unsure whether the arrow in itself would suffice. Finally, Figure 16 shows Plate 142 with the target word *beret*. Here, the developers use a single arrow to narrow the scope of interpretation of the picture.

Figure 14: Stroke weight, color saturation as well as arrow used to make 'hair' salient in a context of face

Figure 15: Framing device and arrow used to make 'isthmus' salient in a context of landmasses and ocean

Figure 16: Arrow used to make 'beret' salient in a context of head, torso and other clothing

On first inspection, Figure 16 appears as a straightforward example of a single repair strategy. But seen in the light of Figure 14, one wonders why stroke weight and color saturation have not been used to make the beret stand out. Or, conversely, whether a simple arrow would have sufficed to repair the picture of hair. There seems to be no system to how the developers have used the four repair strategies identified in the *EOWPVT*. The unsystematic use of several improvement strategies may increase the inference workload required of the child. After all, if the child has recently learnt from Plate 28 *hair* that fat lines specify which part of the picture to pay

attention to, she/he may be confused when fat lines in Plate 142 (where they outline the whole person including the beret) and Plate 63 (where they outline both dentist and patient) have different meaning potentials.

Figures 17-19 below demonstrate extreme cases of what the communicative effects of repair strategies used consistently and inconsistently could be. Whether the inconsistencies observed in the *EOWPVT* in fact pose a problem for the validity of such a test is, of course, a question we may explore empirically. We suggest that, in order to increase the validity of tests that use pictures, we should consider doing so.

Figure 17: Here, three repair devices are used consistently to specify the dentist as worthy of attention

Figure 18: Here, the three repair devices are used as in the *EOWPVT* Plate 63. The result is somewhat ambiguous

Figure 19: Here, the three repair devices compete to specify dentist, patient and lamp as worthy of attention

5. Conclusion

In this article, we have posed that the design of picture prompts in speech and language tests may cause test takers to make different meanings than those intended by the developers of the test. If so, this threatens valid test scoring and results.

In order to investigate this, we first scrutinized the documentation of validity evidence for pictorial aspects of 12 tests currently in common use in Denmark resulting in a categorization of these into four levels of explication. We found that four of the tests provided no evidence for interpretability of picture prompts, three asserted face validity of interpretability, in two tests items had been excluded after statistical analysis after a trial, and only three had some degree of systematic planning in the picture design. None of the developers had employed systematic qualitative analysis of pictures and, insofar as they had been evaluated by experts, none of these were experts in visual communication.

Furthermore, we discussed more or less standardized procedures for generating validity evidence that may lead to better pictorial aspects of tests. Four different approaches were discussed (1) a criterion-keyed approach, (2) an analytic approach, (3) a rational/deductive approach and, finally, (4) item analysis. We concluded that only the rational/deductive approach to test development includes expert evaluations of test items and that all approaches focus on quantitative analysis of item responses in pilot tests. The first part of the study thus concludes that both in the principles for test development overall, and in the actual test materials that we have analyzed, including the test specifications, there is a lack of systematic attention to the meaning potential of pictures used in the tests, which poses a real threat to their validity.

In the second half of the article, we suggested a systematic, qualitative approach to picture design as a possible remedy and proposed an analytic framework based on Kress and van Leeuwen's social semiotic adaptation of Halliday's concept of metafunction as a candidate. We demonstrated the usefulness of such an approach by analyzing reconstructions of test items from the *EOWPVT* and discussed improvement strategies for problematic pictures and whether or not they were effective or potentially misleading.

The implications of the present study are that test developers need to be aware of potential pitfalls of using pictures as prompts in tests. Pictures are not *a priori* meaningful in the sense intended by the developers, especially not for children who are not yet fully visually literate. We demonstrated this by changing the ideational specificity of a drawing of a dentist to show how this could be more or less open to interpretation, the interpersonal specificity to show how it could be more or less engaging, and the textual specificity to the effect that it would likely be interpreted as a picture of a dentist, or of a patient, or of a lamp, respectively. Whether or not these changes would in fact make a difference in a test situation is an empirical question that we suggest pursuing in the future.

Until we know more, we recommend that development standards specifically address issues with the validity of pictures and that test consumers, specifically test administrators, develop a sensitivity to possible pitfalls of using pictures in tests, and develop strategies to secure fair and safe testing. Such sensitivity may be obtained by systematically analyzing pictures included in often used test materials, and by discussing and reflecting within the professional community.

References

AERA, APA & NCME. 2014. *Standards for Educational and Psychological Testing*. Washington D.C.: American Education Research Association.

Bishop, Dorothy V. M. 1977. *Comprehension of Grammar: Normal and Abnormal Development*. (PhD thesis), Oxford University, Oxford.

Bishop, Dorothy V. M. 2010. *TROG-2 Dansk udgave*. Bromma, Sverige: Pearson

Black, Paul. 1998. *Testing: Friend of Foe? Theory and Practice of Assessment and Testing*. London/ New York: RoutledgeFalmer.

Boeriis, Morten. 2009. *Multimodal Socialsemiotik og Levende Billeder*. (PhD thesis), University of Southern Denmark, Odense.

Boeriis, Morten. 2012. Tekstzoom. Om en dynamisk funktionel rangstruktur i visuelle tekster. In Thomas Hestbæk Andersen & Morten Boeriis (eds.) *Nordisk Socialsemiotik. Pædagogiske, Multimodale og Sprogvidenskabelige Landvindinger*. Odense: University Press of Southern Denmark. 131-155.

Brodeur, Mathieu B., Emmanuelle Dionne-Dostie, Tina Montreuil & Martin Lepage. 2010. The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLOS ONE* 5(5), e10773. doi:10.1371/journal.pone.0010773.

Cannard, Christine, Agnès Blaye, Nelly Scheuner & Françoise Bonthoux. 2005. Picture naming in 3- to 8-year-old French children: Methodological considerations for name agreement. *Behavior Research Methods* 37(3). 417-425. doi:10.3758/BF03192710.

Clausen, Marit C. 2014. *LogoFoVa. Logopædisk udredning af Fonologiske Vanskeligheder.* Viborg: Dansk Psykologisk Forlag.

Dahms, Anne-Grethe, Kirsten Loft Bergstrand & Jytte Nakskov. 2005. *Kaj-testen. Spontanbenævnt fonologisk test*. Viborg: Special-pædagogisk forlag.

Dunn, Lloyd M. & Douglas M. Dunn. 2007. *Peabody Picture Vocabulary Test, 4th edn*. London: Pearson Assessment.

Dunn, Lloyd M. & Leota Dunn. 1997. *Peabody Picture Vocabulary Test, 3rd. edn*. London: Pearson Assessment.

Edwards, Susan, Michael Garman, Arthur Hughes, Carolyn Letts & Indra Sinka. 1999. Assessing the comprehension and production of language in young children: an account of the Reynell Developmental Language Scales III. *International Journal of Language & Communication Disorders* 34(2). 151-191.

Edwards, Susan, Carolyn Letts & Indra Sinka,. 2011. *The New Reynell Developmental Language Scales: III (RDS-III)*. London: GL-Assessment.

Edwards, Susan, Carolyn Letts, Indra Sinka, Helle Iben Bylander & Inge Benn Thomsen. 2013. *New Reynell Developmental Language Scales: III - Dansk udgave*. Virum: Hogrefe.

Ege, Bo. 1974. *Sproglydstest*. Herning: Special-pædagogisk forlag.

Ege, Bo. 1987 (revision). *Sproglig test 2*. Herning: Special-Pædagogisk Forlag.

Ege, Bo. 1996. *Sproglydstesten Metha*. Herning: Special-Pædagogisk Forlag.

Ege, Bo. 2003. *Sproglig test 3*. Herning: Special-Pædagogisk Forlag.

Ege, Bo. 2004. *Sproglig test 4*. Herning: Specialpædagogisk Forlag.

Ege, Bo. 2007 (revision). *Sproglig test 1*. Herning: Special-Pædagogisk Forlag.

Geisinger, Kurt F. 2016. Test Evaluation. In Suzanne Lane, Mark R. Raymond & Thomas M. Haladyna (eds.) *Handbook of Test Development*. New York/London: Routledge. 624-638.

Goldstein, Brian, Leah Fabiano & Aquiles Iglesias. 2004. Spontaneous and imitated productions in Spanish-Speaking children with phonological disorders. *Language, Speech, and Hearing Services in Schools* 35(1). 5-15. doi:doi:10.1044/0161-1461(2004/002).

Grønborg, Axel, Ivan Køhler, Jørn Lund & Birthe Høeg Møller. 1987. *DIM - Dansk Impressiv Morfologisk Test*. Herning: Special-Pædagogisk Forlag.

Halliday, Michael & Christian Matthiessen. 2013. *Halliday's Introduction to Functional Grammar*. London & New York: Taylor & Francis.

Haman, Ewa, Magdalena Luniewska & Barbara Pomiechowska. 2015. Designing Cross-linguistic Lexical Tasks (CLTs) for bilingual preschool children. In Sharon Armon-Lotem, Jan de Jong & Natalia Meir (eds.) *Methods for Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*. Bristol, UK: Multilingual Matters. 196-240.

Kane, Michael. 2016. Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In Suzanne Lane, Mark R. Raymond & Thomas M. Haladyna (eds.) *Handbook of Test Development*. New York/London: Routledge. 64-80.

Kaufmann, Birthe, & Bo Ege. 1974. *Sproglydstest*. Herning: Special-Pædagogisk Forlag.

Kress, Gunther & Theo van Leeuwen. 2002. Colour as a semiotic mode: Notes for a grammar of colour. *Visual Communication* 1(3). 343-369.

Kress, Gunther & Theo van Leeuwen. 2006. *Reading Images. The Grammar of Visual Design*. London: Routledge.

Lavesson, Ann, Martin Lövdén & Kristina Hansson. 2018. Development of a language screening instrument for Swedish 4-year-olds. *International Journal of Language & Communication Disorders* 53(3). 605-614. doi:doi:10.1111/1460-6984.12374.

Letts, Carolyn, Susan Edwards, Blanca Schaefer & Indra Sinka. 2013. The New Reynell Developmental Language Scales: Descriptive account and illustrative case study. *Child Language Teaching and Therapy* 29(2). 1-14.

Martin, Nancy A. & Rick Brownell. 2013. *Expressive One-Word Picture Vocabulary Test*. Copenhagen: Dansk Psykologisk Forlag.

Nettelbladt, Ulrika, Christina Samuelsson, Birgitta Sahlén & Marianne Ors. 2008. Språkstörningar hos barn och ungdomar – allmän del. In Lena Hartelius, Ulrika Nettelbladt & Britta Hammarberg (eds.), *Logopedi*. Stockholm: Studentlitteratur.

Pedersen, Elsebeth & Gitte Skyum Kjøge. 2000. *Viborg Materialet*. Viborg: Special-pædagogisk forlag.

Pind, Jörgen, Halla Jónsdóttir, Hjördis Gissurardóttir & Frosti Jónsson. 2000. Icelandic norms for the Snodgrass and Vanderwart (1980) pictures: Name and image agreement, familiarity, and age of acquisition. *Scandinavian Journal of Psychology* 41(1). 41-48. doi:10.1111/1467-9450.00169.

Raykov, Tenko & George A. Marcoulides. 2011. *Introduction to Psychometric Theory*. New York, USA/Hove UK: Taylor & Francis.

Ryan, Joseph J, Shane J. Lopez & Scott W. Sumerall. 2001. Understanding test construction. In William I. Dorfman & Michel Hersen (eds.) *Understanding Psychological Assessment*. New York, Boson, Dordrecht, London, Moscow: Kluwer Academic/Plenum Publishers. 1-15.

Skaggs, Steven. 2017. *FireSigns. A Semiotic Theory for Graphic Design*. Cambridge, MA: The MIT Press.

Slobin, Dan R. 1996. From 'Thought and Language' to 'Thinking for speaking'. In John J. Gumperz & C. Levinson (eds.) *Rethinking Linguistic Relativity*. Cambridge: Cambridge University Press. 70-96.

Snodgrass, Joan Gay & Mary Vanderwart. 1980. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory* 6(2). 174-215.

Stötzner, Andreas. 2003. Signography as a subject in its own right. *Visual Communication* 2(3). 285-302.

Vinberg Hansen, Line. 2017. *Det nemlig smør! Et kvalitativt, konversationsanalytisk, kognitiv-lingvistisk og socialsemiotisk studie af fire børnelogopædiske testsituationer samt testmaterialerne LogoFoVa og Sproglydstesten.* (unpublished master's thesis), University of Southern Denmark, Odense.

Walsh, Maureen. 2003. 'Reading' pictures: what do they reveal? Young children's reading of visual texts. *Reading* 37(3). 123-130. doi:doi:10.1046/j.0034-0472.2003.03703006.x.

Ward, Jamie. 2006. *The Student's Guide to Cognitive Neuroscience*. Hove/New York: Psychology Press.

Wiig, Elisabeth, & Kikki Førsteliin Andersen. 2014. *CELF-4*. Bromma, Sverige: Pearson.