education sciences

Commentary

When Complexity Is Your Friend: Modeling the Complex Problem Space of Vocabulary

Amanda P. Goodwin [1,*], Yaacov Petscher [2], Dan Reynolds [3], Tess Lantos [1], Sara Gould [1] and Jamie Tock [2]

[1] Teaching and Learning, Peabody College, Vanderbilt University, Peabody #230, 230 Appleton Place, Nashville, TN 37203-5721, USA; Tess.lantos@vanderbilt.edu (T.L.); Sara.a.gould@vanderbilt.edu (S.G.)
[2] Florida Center for Reading Research, Florida State University; 2010 Levy Ave., Suite 100, Tallahassee, FL 32310, USA; ypetscher@fcrr.org (Y.P.); Jamie.tock@fsu.edu (J.T.)
[3] Department of Education and School Psychology, John Carroll University; 1 John Carroll Boulevard, University Heights, OH 44118, USA; danielereynolds@gmail.com
* Correspondence: amanda.goodwin@vanderbilt.edu; Tel.: +1-305-710-6257

Received: 1 September 2018; Accepted: 20 September 2018; Published: 15 October 2018

Abstract: The history of vocabulary research has specified a rich and complex construct, resulting in calls for vocabulary research, assessment, and instruction to take into account the complex problem space of vocabulary. At the intersection of vocabulary theory and assessment modeling, this paper suggests a suite of modeling techniques that model the complex structures present in vocabulary data in ways that can build an understanding of vocabulary development and its links to instruction. In particular, we highlight models that can help researchers and practitioners identify and understand construct-relevant and construct-irrelevant aspects of assessing vocabulary knowledge. Drawing on examples from recent research and from our own three-year project to develop a standardized measure of language and vocabulary, we present four types of confirmatory factor analysis (CFA) models: single-factor, correlated-traits, bi-factor, and tri-factor models. We highlight how each of these approaches offers particular insights into the complex problem space of assessing vocabulary in ways that can inform vocabulary assessment, theory, research, and instruction. Examples include identifying construct-relevant general or specific factors like skills or different aspects of word knowledge that could link to instruction while at the same time preventing an overly-narrow focus on construct-irrelevant factors like task-specific or word-specific demands. Implications for theory, research, and practice are discussed.

Keywords: vocabulary; methods; confirmatory factor analysis

1. Introduction

Vocabulary, "the body of words in a given language" [1] is a deceptively complex concept. The term 'vocabulary' can describe the words an individual recognizes (i.e., receptive vocabulary), the words an individual can produce and use (i.e., productive vocabulary), the words that are being learned (i.e., vocabulary learning), as well as the words outside of an individual's understanding. It can describe a broad group of words (i.e., general vocabulary) or a more specific set of words (i.e., academic vocabulary;) or a further specific domain of words (i.e., discipline-specific academic words like those used in science) or even a specific group of words (i.e., weather vocabulary). It can be described from a strength perspective (i.e., considering bilinguals' full range of word knowledge) or from a more deficit-based perspective (i.e., vocabulary gaps). Vocabulary can also describe word-specific knowledge (i.e., definitions, synonyms, antonyms, etc.) or word-general understandings and skills (i.e., metalinguistic awareness, morphological awareness, context clue skills, and word-problem solving). Given this variability, it is no surprise that a century of research has resulted in calls for vocabulary research, assessment, and instruction to take into account the complex problem space of vocabulary (see [2] for an overview). At the core, these conceptual papers call for considering what words, word knowledge, and tasks (i.e., written, oral, decontextualized, and embedded) are being used to explore vocabulary.

The current paper builds on this work of identifying the above complexities to argue that how we model the complexities matters in important ways. This is because innovative modeling techniques allow us to model complex structures that can help researchers and practitioners understand construct-relevant and construct-irrelevant aspects of assessing vocabulary knowledge. These different aspects have important consequences for understanding both how vocabulary develops in children and what is instructionally relevant versus less generative (i.e., construct-irrelevant). An example might be modeling vocabulary consisting of general vocabulary knowledge, specific types of knowledge (e.g., definitional, relational, etc.), and knowledge of either a specific task related to words (e.g., multiple choice vs completing a sentence) or a specific set of words (e.g., a group of words in a story). In terms of understanding a child's vocabulary, modeling vocabulary in this way highlights the bigger picture, allowing one to parse a child's overall vocabulary knowledge from how comfortable they are performing certain cognitive tasks or how well they know a specific word or set of words. When thinking about links to instruction, a teacher might be encouraged to focus on broader construct-relevant components (i.e., building general vocabulary knowledge and doing this through designing activities that build specific types of knowledge like definitions or word relationships), but avoid over-focusing on tasks (e.g., teaching students to successfully figure out multiple choice questions) or on smaller sets of words (e.g., teaching the bold words from a story in the text).

As such, the goal of this paper is to present a modeling framework that can be used to separate out these various components of vocabulary knowledge. We begin by linking to the theory on why studying vocabulary is important. This identifies a rationale for modeling vocabulary in an accurate and detailed manner. Next, we highlight some of the complexities in the vocabulary problem space that such models help us to parse. We then explain how these models fit within the larger statistical landscape. Lastly, we present a set of models to consider when modeling vocabulary. We argue that modeling in this way can build an understanding of vocabulary and sort out construct-relevant structures (i.e., vocabulary relevant) from construct-irrelevant structures (i.e., those more specific to certain cognitive tasks or specific to certain words). For each model, we provide an illustration from the research that shows how such models help us in studying vocabulary. We finish by discussing how such models were pivotal in supporting our work developing a standardized, gamified measure of language and vocabulary for middle schoolers, which we call Monster, PI.

1.1. Why Vocabulary as It Relates to Language and Literacy

Theories of reading and writing agree on one thing: language provides the foundation for literacy [3]. Vocabulary, albeit narrowly, has often been used a proxy for language because both language and vocabulary emphasize the importance of meaning. Many broader language skills (i.e., phonology, syntax, morphology, and semantics) either are directly related to vocabulary (i.e., semantics and morphology) or can be found in models exploring vocabulary. For example, Perfetti's lexical quality hypothesis [4] argues that the properties of the quality of a lexical representation (i.e., what is stored in one's lexicon about a word) include phonological and syntactical information as well as information about meaning. This is why when models of reading highlight the strong role of language, they are usually highlighting the importance of vocabulary in reading as well.

Examples of the role for vocabulary and language in models of reading can be seen across theoretical camps. For example, the Simple View of Reading [5] describes reading as the product of word decoding and linguistic comprehension where linguistic comprehension is defined as "the process by which given lexical (i.e., word) information [vocabulary], sentences and discourses are interpreted" [5] (p. 7). A competing model similarly places language in a central role with Goodman [6] describing reading as "a psycholinguistic guessing game [that] involves an interaction between thought and language" [6] (p. 2). Theorists suggest different reasons regarding how vocabulary knowledge supports reading comprehension. For example, the instrumentalist hypothesis suggests integrating the word meanings themselves is critical whereas the knowledge hypothesis posits that by knowing a word's meaning, one likely knows more about that topic area, leading to improved comprehension [7]. Importantly, for this paper, the theories and research consistently show a relationship between vocabulary and reading [2]. Thus, we believe that it is important to deepen our understanding of this relationship and determine what is relevant and irrelevant when modeling vocabulary. We argue that modeling techniques can help us build this deeper understanding of vocabulary and its relationship to reading.

1.2. Components of the Vocabulary Problem Space

Snow and Kim [8] argue that vocabulary presents a challenging problem space as compared with other literacy learning tasks (particularly decoding) for readers. Similarly, Pearson, Hiebert, and Kamil [2] argue that acknowledging and modeling the complexities of the problem space is necessary to building out understandings that can lead to improvements in both conceptualizations and understandings of links to instruction. In their seminal piece, they write,

> "if we are going to teach it more effectively and if we are going to better understand how it is implicated in reading comprehension, we must first address the vexing question of how we assess vocabulary knowledge and, even more challenging, vocabulary growth. In this essay, we argue that vocabulary assessment is grossly undernourished, both in its theoretical and practical aspects—that it has been driven by tradition, convenience, psychometric standards, and a quest for economy of effort rather than a clear conceptualization of its nature and relation to other aspects of reading expertise, most notably comprehension." [2] (p. 282)

So what does that complicated problem space look like and how is it related to the models we propose? First, it may be helpful to think of vocabulary through dichotomies. These comparisons allow us to hone in on different aspects of vocabulary by explaining differences in the ways we use words. Vocabulary is often contrasted as either written or oral, productive or receptive; additionally, vocabulary knowledge is construed as either broad or deep, word-specific or word-general. These parameters allow us to think about vocabulary as a component of reading and listening, as well as writing and speaking. They also help demonstrate the fact that vocabulary can be a single measurable construct, while also representing a combination of a number of different aspects or components. We discuss a few of these components below.

1.2.1. Cognitive Demand

When thinking about vocabulary, it is helpful to think about the cognitive work involved and its impact on the size and depth of one's vocabulary. Importantly, people have different vocabularies across different cognitive contexts. Generally, receptive vocabulary involves less cognitive demand, meaning that children often score higher on receptive measures compared to productive measures. In other words, most people understand more words through reading and listening than they can produce through speech or writing.

1.2.2. Task Demands

Related, it is also important to consider the demands of vocabulary tasks. Here, Read [9] specified three key concepts for understanding vocabulary assessments: their embeddedness with comprehension, their selection of words to assess, and their tasks' provision of context. Additional test-taking literature highlights differences in item properties such as multiple choice vs. open response, etc. Understanding these three properties of assessments helps inform how our models try to separate construct-relevant variance—that is, students' vocabulary knowledge, from construct-irrelevant variance—that is, how good students are at certain kinds of test-taking tasks.

1.2.3. Breadth vs. Depth

Additionally, when describing vocabulary knowledge, researchers commonly distinguish between breadth and depth. Breadth of vocabulary knowledge refers simply to the quantity of words known, whereas depth of vocabulary knowledge refers to the richness of the understanding of those known words (e.g., [7,10]). While breadth is understood as a numerical quantity (i.e., we can count the number of words that a student knows), depth is thought of as a continuum of understanding, ranging from some recognition of a word to complete understanding of a word's various meanings and how to use it appropriately in a variety of contexts. Depth of vocabulary knowledge includes knowledge of multiple related meanings, including shades of meaning, knowledge of semantically related words including subordinates and superordinates, and the syntactic and pragmatic knowledge of whether and how to use a word in a given context (see later discussion).

1.2.4. Domain of Words

Another thing to consider is which words to assess and what performance on a certain set of words means. There are more than 200,000 words that readers might encounter in academic texts [11]. Recent work indicates there are a core group of words present in these texts (i.e., 2451 morphological families making up 58% of the 19,500 most frequent words in text [12] and we also know there are a larger set of academic words that students tend to encounter in their school learning [13–15]. We further know that words are learned in networks, including conceptual groups [16] and in morphological families [17]. This suggests that overall word knowledge likely matters more than knowledge of a specific word or set of words, unless that set of words represents a larger instructionally relevant principle (such as a concept or skill like learning words in morphological families).

1.2.5. Word-Specific vs. Word-General

Another distinction in vocabulary knowledge is often made between word-specific knowledge and word-general knowledge [18]. Word-specific knowledge includes breadth and depth of knowledge of individual word meanings. Word-general knowledge, on the other hand, involves metalinguistic knowledge about words and their meanings. One component of word-general knowledge is morphological awareness [19]. Morphology is the system by which smaller meaningful units, such as affixes and roots, are combined to form complex words. These units of meaning contribute to both the overall meaning and the syntactic function of the word, making them an important word-generative understanding. Another component of word-general knowledge is students' skill in using context to provide information about word meanings, or their contextual sensitivity [19].

1.2.6. Word Characteristics

Yet another distinction relates to how not all words are created equal in terms of word learning. Vocabulary assessments need to consider the characteristics of words and how those characteristics relate to word learning. We identify these characteristics below in Table 1 and refer readers to a discussion of these characteristics [12]. The point, though, of emphasizing word characteristics is that these differences in words make it even more important to get at construct-relevant and construct-irrelevant variance.

Table 1. Key Word Characteristics.

Word Length	Meaning (Semantics and Syntax)	Frequency/Familiarity	Orthographic Representation
	• Polysemy	• Word family; root(s)	• Regularity of grapheme to phoneme correspondence
• # Letters	• Conceptual complexity	• Concreteness	• Transparency at syllable and morpheme levels
• # Syllables	• Domain specificity	• Age of acquisition	• Irregular spellings
• # Morphemes	• Syntactic/grammatical roles	• Morphemic composition/transparency	
		• Word origin	

1.2.7. Aspects of Word Knowledge

Perhaps the most important consideration within the larger problem space related to studying vocabulary is considering what is known about word knowledge, and this relates to the discussion of depth above. Here, we highlight Nagy and Scott's [20] principles of (1) incrementality, which specifies that word knowledge is not a dichotomous known–unknown state but a continuous spectrum of understanding; (2) polysemy, which notes that words often have multiple meanings; (3) multidimensionality, suggesting that word knowledge includes different kinds of knowledge such as denotation, connotation, or collocation; (4) interrelatedness, that is, word knowledge is linked to conceptually related understandings; and (5) heterogeneity indicating that a word's meaning is dependent on its function and structure. The reason these categories are so important is that these five properties of word knowledge inform the types of factors that might be modeled. We go into further detail regarding each of these aspects of word knowledge later.

A common-sense idea holds that word knowledge is incremental. A low amount of knowledge might involve merely recognizing a familiar word whereas much more knowledge is required to have a subtle and nuanced understanding of a word's multiple uses. In fact, as researchers have noted, incremental knowledge of word meaning is similar to vocabulary depth (explained previously), which has been shown to be crucial from early childhood through college age and adulthood [21,22]. Beyond incremental knowledge, words also have multiple meanings—that is, they are polysemous. In fact, common words such as "set" and "run" have hundreds of meanings [1], ranging from everyday to quite specialized. Notably, Logan and Kieffer [23] found that seventh-grade students' knowledge of the particular academic meanings of words (that is, the polysemity of their vocabulary) uniquely predicted their comprehension, even when controlling for vocabulary breadth, awareness of everyday meanings of words, and decoding skills.

The third principle, multidimensionality, links to the idea of Perfetti's lexical quality hypothesis [4]: word knowledge is not just the semantics, but the orthographic, phonological, morphological, syntactic, and pragmatic aspects of its practical use. Nagy and Scott [20] point out that one student might recognize a word's orthographic form but be unable to use it in speech, while others might use a word in the proper academic register but misuse its denotation. This leads to the principle of interrelatedness, which notes that words exist in conceptual relation to other words. That is, "a person who already knows the words hot, cold, and cool has already acquired some of the components of the word *warm*" [20] (p. 272). Fitzgerald, Elmore, Kung, and Stenner [16] argue that, especially in science learning, the development of complex networks of interrelated vocabulary knowledge co-develops with children's understandings of complex concepts about the world around them. This is important because the goal of vocabulary growth is not an end, but a means of supporting students in communicating effectively with a myriad of audiences. The final principle, heterogeneity, posits that word knowledge includes knowledge of functional use. Here, word knowledge does not exist in isolation but as multifarious sources that can be used as tools to enable communication, with particular features for academic communication [13]. These principles show that a word's semantic knowledge is far more complex than knowledge of a single definition would imply and inform our view of possible construct-relevant factors to model within vocabulary.

Overall, the problem space related to vocabulary is clearly complex. Innovative vocabulary assessments that attend to the complexities of the vocabulary problem space are needed, as well as modeling techniques that can use these complexities as an asset rather than a hindrance. Next, we suggest some possible frameworks to use that take advantage of these complexities when modeling. These techniques can be used to unravel construct-relevant from construct-irrelevant variance. Note that the important part of these frameworks is that they attend to multiple complexities within the vocabulary problem space.

2. Potential Frameworks

As an introduction to our presentation of models, it is important to revisit a few *hows* and *whys* of statistical modeling. Statistical modeling maintains goals of prediction and explanation [24]. Within the explanatory mechanism, the latent variable modeling approach of confirmatory factor analysis (CFA) estimates how well a hypothesized model can fit, or partition well, the variance–covariance matrix to a set of user-specified latent factors indicated by manifest variables. Here, the researcher can compare models with different hypothesized structures, and the extent to which a specified model fits well according to combinations of incremental, comparative, or absolute fit indexes allows a researcher to begin identification of construct-relevant or construct-irrelevant sources of variance [25].

Consider the example of reading comprehension. In many assessment situations, students are administered multiple passages to read followed by a set of comprehension questions. Researchers will often simply sum the scores across the items without considering that the items have important contextual factors. One contextual factor that is frequently ignored is that items exist within passages; passages can be expository or narrative. Ignoring the potential influence of expository or narrative text on students' performance might limit our understanding of individual differences in students' reading comprehension and impact instructional implications. Another contextual factor is that across a set of passages there may be items within each passage that are designed to measure instructionally relevant skills, such as the author's purpose, selecting the main idea, or vocabulary. Here too, simply summing the items across the passages to obtain a global reading comprehension score not only limits the utility of the score but could provide misinformation about a student's ability.

CFA affords a number of modeling choices that surpass the simple summing of scores, and these modeling choices maintain inherent flexibility to include, among other attributes, single or multiple latent variables, simple or complex loading solutions, and correlated or uncorrelated constructs. One of the most well-known expressions of a CFA is a single-factor, or unidimensional, model whereby one latent construct is specified to be indicated by one or more measured variables. When more than one latent construct is specified, the CFA is known as a multidimensional or correlated trait model. Within this multidimensional model, each of the two or more constructs are indicated by respective, or sometimes cross-loading, measured variables and the constructs are typically estimated to correlate with each other. More complex CFAs can be readily specified, but for the purposes of this paper we focus on four types: (1) the single-factor model, (2) the correlated trait model, (3) the bi-factor model, and (4) the tri-factor model.

Beyond single factor and correlated trait models, a bi-factor model [26] hypothesizes that the measured variables simultaneously indicate two latent variables: A global construct designed to measure construct-relevant variance, and one or two or more latent constructs (i.e., sometimes referred to as *specific constructs*) that measure structures within and uniquely separable from the global construct. These specific constructs can be construct-relevant (e.g., skills like author's purpose, selecting the main idea, etc) or construct-irrelevant (e.g., multiple choice vs. cloze responses). The specific construct(s) are indicated by a subset of items whereas the global construct is indicated by items across the specific constructs. For example, when students read multiple reading comprehension passages, the item-level responses to questions may include information that loads simultaneously on a global latent variable of reading comprehension as well as a specific latent variable that is specific to the passage to which the item belongs [27]. An example of construct-relevant variance might be the global latent variable of reading comprehension and an example of construct-irrelevant variance would be the latent variables associated with passage effects. The global reading comprehension factor, in this case, would be malleable for teachers to instruct on (i.e., construct-relevant variance), and the specific passage effects would be nuisance constructs (i.e., construct-irrelevant variance [28]). Latent factors in the bi-factor model are typically set to be uncorrelated as the theory of bi-factor models is that the covariance among measured variables is not best explained by latent correlations for specific constructs but instead represented by a global construct. An example of a bi-factor model in language research is Goodwin et al. [29] who found that a bi-factor model was best for representing morphological knowledge data.

A tri-factor model is an extension of a bi-factor where measured variables would simultaneously indicate three latent variables. This model has not been frequently used within educational research; however, psychology researchers have used it to unpack variance attributions to theoretical constructs, maternal raters, and paternal raters [30]. With these exemplary CFA model-types in mind, our framework building approach described below is focused on identifying construct-relevant and construct-irrelevant variance with the idea of modeling vocabulary in ways that lead to deeper understandings of vocabulary development and links to instruction.

3. Proposed Models

3.1. Model 1: Single Factor Model

We begin by thinking about how to measure what students know about words in general. We first consider vocabulary as a single, albeit complex, construct. Here, we draw on the consistent findings that vocabulary is related to reading comprehension (i.e., usually correlations between 0.6 and 0.8 between reading comprehension and vocabulary [2]) and in general, people who have strong vocabularies tend to know a lot about words. The fact that a reader does not know a specific low-frequency word or even piece of word knowledge (i.e., a specific meaning) is unlikely to affect reading in general [31]. The single factor model (i.e., unidimensional model, see Figure 1) uses multiple measures of vocabulary, but conceptualizes these measures as overlapping. In other words, what is measured by one vocabulary measure might be slightly different from the other vocabulary measure, but both measures assess a broader vocabulary construct (i.e., what students know about words). Here, a single construct is modeled and, when presenting with significant variance or loadings, would be assumed to have construct-relevant variance.

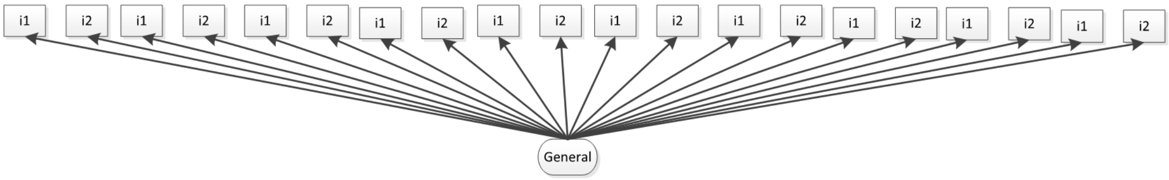

Figure 1. Model 1: Single Factor Model.

Conceptualizing vocabulary as unidimensional tends to be a default perspective in terms of modeling vocabulary for much of educational research (see [2,19] for explanations). It is also one of the first steps when exploring whether modeling additional complexities makes sense within a researcher's vocabulary data. For example, Kieffer and Lesaux [19] explored whether vocabulary was best modeled as a single (i.e., unidimensional) factor or a multidimensional construct (i.e., correlated trait model). Here, sixth graders were assessed on thirteen reading-based vocabulary measures with the goal of better understanding vocabulary knowledge. These measures assessed knowledge of synonyms, multiple meanings, context clues, and morphological knowledge, with all assessments assessing different words. Confirmatory factor analysis suggested that a single factor model for this data was not as good of a fit to the data as a multidimensional model (i.e., correlated trait model), providing evidence of the importance of modeling these structures within vocabulary data.

3.2. Model 2: Correlated Trait Model

A primary finding of the research reviewed above is that there are different aspects of word knowledge that are key to the development of children's vocabulary and which likely link to how words are instructed. Correlated trait models (see Figure 2) can model these different aspects as unique and related. Here, the key principle is to measure multiple aspects of word knowledge (e.g., definitional, relational, and incremental), with the expectation that the measures will group together as a means of conveying construct-relevant structures. The way in which the measures group together inform how a person's lexicon (i.e., vocabulary) develops. With that said, the components usually are hypothesized as meaningful to the larger construct and therefore, again, the correlated trait model conceptualizes components as construct-relevant, largely ignoring construct-irrelevant structures.

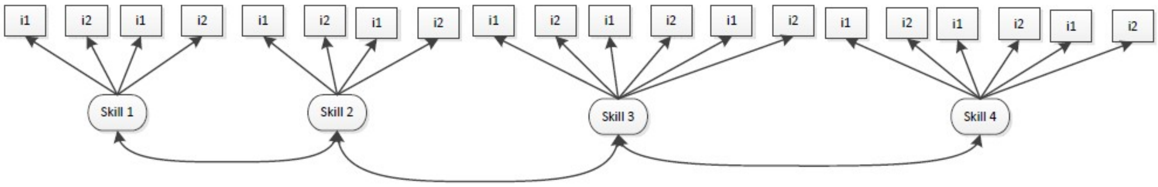

Figure 2. Model 2: Correlated Trait Model.

An example of this can be found in Kieffer and Lesaux's study [19] described previously. Use of the correlated trait model showed the vocabulary data best fit this multidimensional model and showed how the different measures of word knowledge might come together within one's lexicon (i.e., aspects of word knowledge that one might know about words). The results indicated the best fitting model consisted of three highly related, but distinct dimensions, which the authors termed breadth, contextual sensitivity, and morphological awareness. Breadth was made up of assessments where students had to provide synonyms, multiple meanings, and semantic associations for given words. Contextual sensitivity was made up of assessments of students' success at using context clues to determine word meanings. Morphological awareness was made up of real-world decomposition tasks and nonword suffix tasks. This model held for both fluent English students and dual language learners who spoke a language other than English at home, suggesting its robustness and importance in future vocabulary modeling. In this study, the correlated trait model was able to identify specific aspects of vocabulary knowledge (i.e., groupings of vocabulary tasks) that helped the researchers understand how vocabulary knowledge developed for fluent English students as compared to students who were developing multiple languages (i.e., spoke a language other than English at home). Linking to the larger vocabulary literature, the three dimensions of vocabulary knowledge modeled appear to represent different aspects of word knowledge, skills (context clue use), and linguistic awareness (morphological awareness) that could be used to figure out word meanings and build vocabulary.

The benefit of using correlated trait models when modeling vocabulary is that researchers can begin to identify structures present in vocabulary knowledge, or important aspects of word knowledge to model. Here, the individual factors are assumed to maintain construct–relevant variance. Additionally, in correlated-traits models, the constructs are estimated to correlate with each other, and often do to a high level. Those high correlations could be hiding a different source of variance that is not modeled which could be better captured by a general factor (see description of next model). For example, in the study described previously [19], the correlations between vocabulary breadth and contextual sensitivity were high (r = 0.85). This correlation may be indicative of variance that could be captured by a general factor of vocabulary knowledge. This leads us to our next model.

3.3. Model 3: Bi-Factor Model

While dimensionality findings establish related but distinct dimensions of vocabulary knowledge (i.e., multiple aspects of word knowledge), these models do not address how the factors overlap and how they might be separate from the larger construct. Here, we argue that while these constructs are indeed distinct, the relationships between them represent meaningful overlap such that it is best modeled as a higher level construct. In other words, the measured variables are best represented by a bi-factor model (see Figure 3) with an overarching global construct that stems from the overlap of all items as well as specific latent constructs that measure underlying structures within and uniquely separable from this global construct.

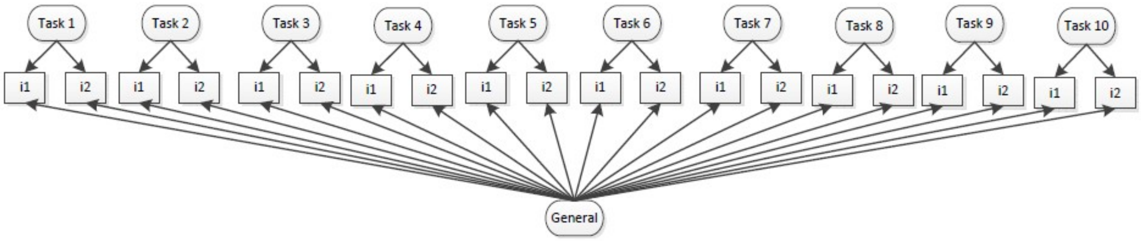

Figure 3. Model 3: Bi-factor Model.

Two recent publications show the importance and power of using bi-factor models when modeling vocabulary. First, when modeling morphological knowledge, Goodwin, Petscher, Carlisle, and Mitchell [29] showed that performance on seven morphological knowledge tasks for middle schoolers was best represented by a bi-factor model consisting of a general factor of morphological knowledge and specific factors representing differences in how tasks assessed different facets of morphological knowledge. For example, some tasks assessed morphological knowledge related to decoding while another assessed morphological knowledge as applied to determining word meanings. While tasks have been mentioned in the literature review as representing construct-irrelevant variance, in this study, variance related to tasks is construct-relevant because the task conveys a skill that is potentially instructionally relevant (e.g., applying morphological understandings to figuring out a word's meaning) rather than a cognitive action (e.g., choosing amongst answers in a multiple choice task) that is unrelated to the larger construct. These researchers showed different predictive relationships to larger literacy constructs between the general and specific factors identified. For example, the general factor and the specific factor of morphological meaning processing were positively related to standardized reading comprehension and vocabulary, whereas the specific factor of generating morphologically related words was only positively associated with standardized vocabulary knowledge.

Another example of a bi-factor model is Kieffer, Petscher, Proctor, and Silverman's [32] analysis of language skills. Here, two different studies presented in this single paper showed multiple supports for conceptualizing language as a bi-factor model. In these studies, language was made up of a general factor and specific factors of morphological awareness and vocabulary and in one study, syntax. The main idea related to assessing vocabulary is that assessments of vocabulary likely tap into general vocabulary, which is itself important to language and literacy performance. Assessments of vocabulary similarly tap knowledge of specific aspects of vocabulary which themselves represent unique pieces of knowledge or skills that are likely instructionally relevant. In the two studies described, the general and specific factors were conceptualized as construct-relevant, but bi-factor models can also identify construct-irrelevant variance (i.e., tasks) depending on design. For example, in the Goodwin et al. study [29], the specific factors related to tasks represented meaningful ways that morphological skills were applied to different literacy tasks. Alternatively, the tasks could have been related to specific cognitive skills rather than morphological skills, which would allow for modeling the construct-irrelevant variance discussed. The challenge is to have enough measures to capture all these structures, and for that, we turn to our next model, the tri-factor model.

3.4. Model 4: Tri-Factor Model

The tri-factor model (see Figure 4) models additional structures; it extends the bi-factor model to include three latent variables (a general factor, specific factors related to one structure like the aspects of word knowledge described above, and specific factors related to a different structure like task differences described above). As mentioned in our discussion of these statistical models, tri-factor models are rarely used in educational research. Still, given the layers of complexity involved in assessing vocabulary, they may be useful when researchers have enough assessments to model various structures. Factors, then, can be modeled as latent variables that are potentially construct-relevant or construct-irrelevant. Based on the vocabulary literature, a way that tri-factor models may be useful when modeling vocabulary is that the first latent variable type would include tasks. This conceptualization of tasks is, at the lowest level, representing different cognitive actions (e.g., multiple choice vs. sentence completion) that are construct-irrelevant such that an educator would not want to instruct how to complete the specific tasks because the latent skills, not the tasks, are what matters. In other words, the educator would not want to instruct students how to complete a multiple choice assessment, instead teaching the content or skills being assessed. The second latent variable type includes skills or components of knowledge. These are the components that other models have found (e.g., building definitions, understanding interrelatedness, using context clues, and unraveling heterogeneity) that are construct-relevant and instructionally relevant such that a teacher would want to instruct these larger skills. Here, for example, an educator might teach how to use units of meaning like root words or affixes to determine the meaning of an unfamiliar word (i.e., the skill of applying principles of morphology to determining meaning of words). The third latent variable type is the global trait, here vocabulary knowledge, which represents overall ability on the larger construct. An educator or researcher might use this information to determine overall performance and, in the case of vocabulary, evaluate whether building vocabulary is something that needs to be targeted intensely for example if a student scored low or whether it could be used to support other literacy skills if the student scored well. To the best of our knowledge, work has not been published in this area, but the problem space of vocabulary suggests it might be a helpful modeling tool. Initial analyses of our morphology data from our three years of development of a standardized gamified language assessment (Monster, PI to be discussed later) indicates this structure fits best. Here, 15 piloted tasks are best fit by a general factor of morphological knowledge (construct-relevant), four skills representing specific factors or specific components of morphological knowledge (construct-relevant), and specific factors related to each task used to measure morphological knowledge (construct-irrelevant). Because we had multiple tasks assessing each skill, the tri-factor model allowed us to separate the construct-relevant variance related to skills from the construct-irrelevant variance related to the cognitive skills needed to do each task. This clarifies structures that are usually ignored without such modeling techniques.

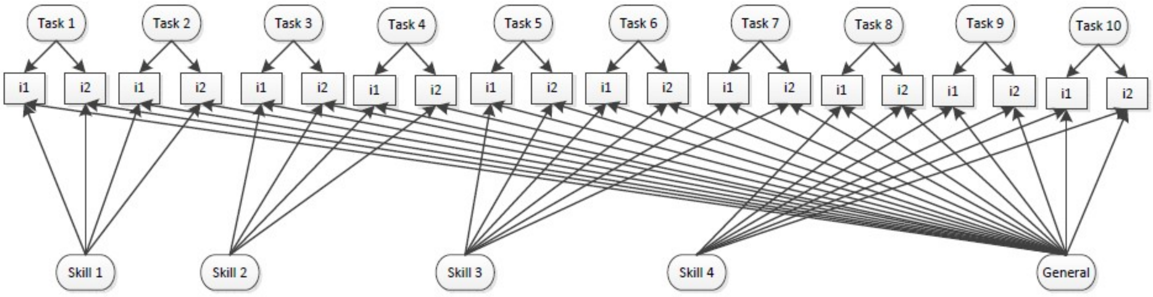

Figure 4. Model 4: Tri-factor Model.

The potential power of understanding the relationships between tasks and skills is that it offers teachers concrete examples of what a particular skill looks like. More importantly, it has the potential to show teachers multiple ways to support students' vocabulary learning. For example, an assessment which gives teachers information about each of the ten tasks might be hard to translate into instruction, as the tasks themselves might not translate into instructional activities. Equally challenging to use might be an assessment which gives teachers information about a single dimension of vocabulary knowledge (e.g., morphology); teachers might struggle to break that large problem space into instructional activities. Grouping the tasks into clearly articulated subskills, however, might offer both a stronger rationale for teaching these skills as well as presenting teachers with multiple examples of activities and assessments that tap those skills. These examples could be starting points for teachers who are themselves learning about the complexity of vocabulary knowledge.

Another example of how the tri-factor model could help sort construct-relevant from construct-irrelevant variance includes models that allow for specifying specific factors for the specific words studied. Although dimensionality related to individual words has been shown in work using other models [17,33], tri-factor models allow for constructing three latent variables: one representing general knowledge, one representing skills, and one representing words. Here, at the lowest level, the latent factors representing each word reflects knowledge of each specific word, which tends to be construct-irrelevant because knowledge of a single word or even a single set of words is unlikely to support larger literacy efforts. In other words, knowledge of weather vocabulary only supports literacy endeavors that include those specific words in texts, and, therefore, the goal is to teach large networks of words rather than word-by-word. The second latent variable type includes skills or components of word knowledge, similar to that described above. Here, performance is construct-relevant because this involves larger principles of word instruction such as highlighting the polysemous nature of words. The third latent variable represents general vocabulary knowledge, which again provides educators with an overall understanding of the students' understandings of words in general. Our work in developing the Monster, PI language assessment, highlighted later, indicates the tri-factor model was the best fitting model for our vocabulary data. In our data, the skills represented performance in definition, word relation, verbal analogy, and polysemy tasks, and these skills were separable from performance on each specific word. Linking to theory and instruction, it is likely that general vocabulary knowledge and students' different aspects of word knowledge are important factors in their language and literacy efforts (i.e., construct-relevant), but that their performance on a specific word is less relevant and less in need of instruction (i.e., construct-irrelevant).

3.5. Development of a Standardized Language Assessment

The models described previously were critical in our work developing a standardized, gamified middle school language assessment, which we call Monster, PI. This assessment assesses student understandings of written language, specifically of morphology, vocabulary, and syntax. Because it is a computer adaptive test (CAT), students start by taking a set of items and then take harder or easier items depending on their performance. How these items are delivered is based on which of the structures described above fits bests. Items continue to be delivered until a stable score is determined that represents a student's abilities. The items are embedded in a game format as 'puzzles' that students solve to earn clues to capture a monster who is wreaking havoc on the city. Crucial here is that we used these models to figure out the best structures to represent our constructs and data, resulting in a valid and reliable assessment.

We began by considering language and each construct within language (i.e., morphology, vocabulary, and syntax) as potentially multidimensional. We then identified different sources of construct-relevant and construct-irrelevant variance to consider. Informed by the models above, we considered various tasks and words in which skills could be embedded such that if the structures fit accordingly, we could separate out skill performance from task performance and/or performance on specific words. At the same time, we could also look across performance on all skills to identify general construct performance, allowing Monster, PI to convey general written language ability as well as ability in morphology, vocabulary, and syntax.

Identifying and choosing tasks and then matching them with skills was challenging, but because our work was informed by the models above, we designed and assessed multiple measures for each word, skill, and construct allowing the complexity of the problem space of language and vocabulary to serve as supports in building our assessment. Our hypothesized models often had to be adjusted based on our combination of theory and statistics, and this family of models ultimately provided a structure to our data that we could then convey to teachers in a meaningful way.

An advantage to our final model was the ability to sort out construct-relevant from construct-irrelevant data as both our morphology and vocabulary models ultimately fit best as tri-factor models where the construct-relevant latents were the general factors (i.e., general vocabulary and general morphological knowledge) and the skills being assessed. For vocabulary, skills included the ability to determine (1) definitions, (2) word relations like antonyms and synonyms, (3) polysemous meanings, and (4) verbal analogies. For morphology, skills identified included (1) identification of units of meaning, (2) use of suffixes, (3) application to meaning, and (4) reading and spelling of morphologically complex words. Scores showing performance in these skills are construct relevant because an educator could design instruction around each of these skills. The construct-irrelevant latents represented specific knowledge of words (i.e., for vocabulary) or tasks (i.e., for morphology). As mentioned, these are construct-irrelevant because teaching a single word or set of words would be unlikely to build general vocabulary and similarly, teaching a certain task like multiple choice or fill in the blank would not build the desired understandings related to vocabulary. For each model, we initially hypothesized additional tasks that were ultimately not retained because they did not explain additional construct-relevant variance. Lessons here indicate that development of vocabulary assessments must include multiple tasks and then must consider whether each task contributes meaningfully to the larger model. Additionally, for both constructs, we originally hypothesized an additional skill which ultimately did not fit our data, suggesting that these models can be helpful in identifying potentially relevant skills for instruction and eliminating others. These models have supported us in developing a reliable and valid standardized, gamified, computer adaptive assessment of language. This assessment will serve both researchers and practitioners in developing understandings of what students know and considering possible instructional interventions for better supporting developing readers.

4. Discussion

Empirical research and theory development in vocabulary paint a complex picture of vocabulary. Researchers continue to grapple with how to best model complexities related to vocabulary. In this paper, we have presented a family of models that can capitalize on many of the complexities identified within the larger vocabulary literature to model structures that build an understanding of children's vocabulary development and links to instruction. Specifically, we argue that these structures can help researchers model construct-relevant variance vs. construct-irrelevant variance. Such distinctions are critical to obtaining accurate understandings that can move the field forward.

The various modeling strategies suggested offer different insights for vocabulary research. Single factor vs multidimensional models can help researchers understand their own data in light of what the larger field is suggesting: that vocabulary is multidimensional in various ways. There is a use for single factor models, though. A researcher may be interested in one aspect of vocabulary or may find that some of the dichotomies suggested within the literature (e.g., breadth vs. depth) do not stand when tested in these models or may even vary for learners at different ages or who are considering different sets of words. As such, single factor models can serve to highlight similarities and overlap within the larger vocabulary landscape. Multidimensional models, in contrast, can serve to highlight potentially important similarities and differences. Correlated trait models identify unique, but often correlated factors. Bi-factor models offer a larger understanding, exploring structures for these more specific factors but also modeling a larger general factor. Bi-factor and tri-factor models help us model additional structures that can be important to moving the field forward. Here, distinctions between construct-relevant and construct-irrelevant variance take center stage. These models allow for modeling a general factor and specific instructionally relevant skills while at the same time separating out noise that may be distracting from these more important understandings. This noise might be related to how students perform on a particular task or a particular word or set of words.

Limitations

One thing to keep in mind is that these models are just a starting point, and also have their challenges as well. First, these models require extensive data collection (e.g., 13 different measures of vocabulary [19]) and detailed assessment design (e.g., our work with Monster, PI, which examined and tested more than 20 tasks over three years). Additionally, replication studies are important as models are likely to vary when considering student differences such as those related to language background or age. While some of the studies contained students from diverse language backgrounds (e.g., [17,19]), consistent attention to the nuances of vocabulary remains essential, including how students' local and vernacular English vocabulary intersects with the principles of academic English. In addition, all of the studies described here focus on students' vocabulary learning at the crucial ages of upper elementary through the end of middle school (i.e., grades 4–8). How might these models translate to learners across the developmental spectrum? Furthermore, while these studies offer exciting insights into the complex problem space of vocabulary, it can be difficult to translate the insights of these models into instructional practice. Developing teacher-friendly ways to discuss the implications of these models could be a way for this research to better impact students.

Overall, the models described here are a sampling of choices one may use when exploring the richness of complexity related to vocabulary. Beyond these sets of models, other emerging analytics assist researchers with simultaneously modeling the relation between person-level attributes and word-level attributes in understanding individual differences related to word-level performance [17,34,35]. These models are known by a variety of names including generalized linear mixed models, cross-classified models, and explanatory item response models. The framework of these methods involves parsing and explaining variance in item-level responses that are due to individual differences and that are due to item differences [36]. Such models take into account the fact that different readers learn different words differently and hence might be useful next steps. An additional step is to model interactions with instruction [17] as word learning is often dependent on instruction.

5. Conclusions

The models suggested in this paper mark an important first step in modeling complexities related to vocabulary, including construct-relevant vs. construct-irrelevant variance. Such models can be helpful when designing a vocabulary assessment and when modeling vocabulary more broadly. It is our hope that illustrating and comparing these models might inform future study designs that can capitalize rather than struggle with the complexity of vocabulary. We hope this discussion will support others in selecting the appropriate analytical models to enrich theory and practice.

Funding: This research was funded by the Institute of Educational Sciences, grant number R305A150199.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vocabulary. In Oxford Dictionaries. 2018. Available online: https://en.oxforddictionaries.com/definition/vocabulary (accessed on 1 September 2018).
2. Pearson, P.D.; Hiebert, E.H.; Kamil, M.L. Vocabulary Assessment: What We Know and What We Need to Learn. *Read. Res. Q.* **2007**, *42*, 282–296. [CrossRef]

3. Dickinson, D.K.; Golinkoff, R.M.; Hirsh-Pasek, K. Speaking out for Language: Why Language is Central to Reading Development. *Educ. Res.* **2010**, *39*, 305–310. [CrossRef]
4. Perfetti, C. Reading Ability: Lexical Quality to Comprehension. *Sci. Stud. Read.* **2007**, *11*, 357–383. [CrossRef]
5. Gough, P.B.; Tunmer, W.E. Decoding, Reading, and Reading Disability. *Rem. Spec. Educ.* **1986**, *7*, 6–10. [CrossRef]
6. Goodman, Y.M. *Reading Miscue Inventory: Alternative Procedures*; Richard C. Owen Publishers: New York, NY, USA, 1987.
7. Anderson, R.C.; Freebody, P. Vocabulary knowledge. In *Comprehension and Teaching: Research Reviews*; Guthrie, J.T., Ed.; International Reading Association: Newark, DE, USA, 1981; pp. 77–117.
8. Snow, C.E.; Kim, Y.-S. Large Problem Spaces: The Challenge of Vocabulary for English Language Learners. In *Vocabulary Acquisition: Implications for Reading Comprehension*; Wagner, R.K., Muse, A.E., Tannenbaum, K.R., Eds.; Guilford Press: New York, NY, USA, 2007; pp. 123–139.
9. Read, J. *Assessing Vocabulary*; Cambridge University Press: Cambridge, UK, 2000.
10. Stahl, S.A.; Nagy, W.E. *Teaching Word Meanings*; Routledge: Hoboken, NJ, USA, 2006.
11. Nagy, W.E.; Anderson, R.C. How Many Words are there in Printed School English? *Read. Res. Q.* **1984**, *19*, 304–330. [CrossRef]
12. Hiebert, E.H.; Goodwin, A.P.; Cervetti, G.N. Core Vocabulary: Its Morphological Content and Presence in Exemplar Texts. *Read. Res. Q.* **2018**, *53*, 29–49. [CrossRef]
13. Nagy, W.; Townsend, D. Words as Tools: Learning Academic Vocabulary as Language Acquisition. *Read. Res. Q.* **2012**, *47*, 91–108. [CrossRef]
14. Snow, C.E.; Lawrence, J.F.; White, C. Generating Knowledge of Academic Language among Urban Middle School Students. *J. Res. Educ. Eff.* **2009**, *2*, 325–344. [CrossRef]
15. Uccelli, P.; Galloway, E.P.; Barr, C.D.; Meneses, A.; Dobbs, C.L. Beyond Vocabulary: Exploring Cross-Disciplinary Academic-Language Proficiency and Its Association with Reading Comprehension. *Read. Res. Q.* **2015**, *50*, 337–356. [CrossRef]
16. Fitzgerald, W.J.; Elmore, J.; Kung, M.; Stenner, A.J. The Conceptual Complexity of Vocabulary in Elementary-Grades Core Science Program Textbooks. *Read. Res. Q.* **2017**, *52*, 417–442. [CrossRef]
17. Goodwin, A.P.; Cho, S.J. Unraveling Vocabulary Learning: Reader and Item-Level Predictors of Vocabulary Learning within Comprehension Instruction for Fifth and Sixth Graders. *Sci. Stud. Read.* **2016**, *20*, 490–514. [CrossRef]
18. Nagy, W.E. Metalinguistic Awareness and the Vocabulary-Comprehension Connection. In *Vocabulary Acquisition: Implications for Reading Comprehension*; Wagner, R.K., Muse, A.E., Tannenbaum, K.R., Eds.; Guilford: New York, NY, USA, 2007; pp. 52–77.
19. Kieffer, M.J.; Lesaux, N.K. Direct and Indirect Roles of Morphological Awareness in the English Reading Comprehension of Native English, Spanish, Filipino, and Vietnamese Speakers. *Lang. Learn.* **2012**, *62*, 1170–1204. [CrossRef]
20. Nagy, W.E.; Scott, J.A. Vocabulary Processes. *Handb. Read. Res.* **2000**, *3*, 269–284.
21. Binder, K.S.; Cote, N.G.; Lee, C.; Bessette, E.; Vu, H. Beyond Breadth: The Contributions of Vocabulary Depth to Reading Comprehension among Skilled Readers. *J. Res. Read.* **2017**, *40*, 333–343. [CrossRef] [PubMed]
22. Hadley, E.B.; Dickinson, D.K. Measuring Young Children's Word Knowledge: A Conceptual Review. *J. Early Child. Lit.* **2018**, in press.
23. Logan, J.K.; Kieffer, M.J. Evaluating the Role of Polysemous Word Knowledge in Reading Comprehension among Bilingual Adolescents. *Read. Writ.* **2017**, *30*, 1687–1704. [CrossRef]
24. Shmueli, G. To Explain or to Predict? *Stat. Sci.* **2010**, *25*, 289–310. [CrossRef]
25. Morin, A.J.; Arens, A.K.; Marsh, H.W. A Bifactor Exploratory Structural Equation Modeling Framework for the Identification of Distinct Sources of Construct-Relevant Psychometric Multidimensionality. *Struct. Equ. Model.-Multidiscip. J.* **2016**, *23*, 116–139. [CrossRef]
26. Reise, S.P. The rediscovery of Bifactor Measurement Models. *Multivar. Behav. Res.* **2012**, *47*, 667–696. [CrossRef] [PubMed]
27. Petscher, Y.; Foorman, B.R.; Truckenmiller, A.J. The Impact of Item Dependency on the Efficiency of Testing and Reliability of Student Scores From a Computer Adaptive Assessment of Reading Comprehension. *J. Res. Educ. Eff.* **2017**, *10*, 408–423. [CrossRef]

28. Wainer, H.; Bradlow, E.T.; Wang, X. *Testlet Response Theory and Its Applications*; Cambridge University Press: Cambridge, UK, 2007.
29. Goodwin, A.P.; Petscher, Y.; Carlisle, J.F.; Mitchell, A.M. Exploring the Dimensionality of Morphological Knowledge for Adolescent Readers. *J. Res. Read.* **2017**, *40*, 91–117. [CrossRef] [PubMed]
30. Bauer, D.J.; Howard, A.L.; Baldasaro, R.E.; Curran, P.J.; Hussong, A.M.; Chassin, L.; Zucker, R.A. A Trifactor Model for Integrating Ratings across Multiple Informants. *Psychol. Methods* **2013**, *18*, 475–493. [CrossRef] [PubMed]
31. Perfetti, C.A.; Hart, L. The Lexical Bases of Comprehension Skill. In *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity*; Gorfien, D.S., Ed.; American Psychological Association: Washington, DC, USA, 2001; pp. 67–86.
32. Kieffer, M.J.; Petscher, Y.; Proctor, C.P.; Silverman, R.D. Is the Whole Greater than the Sum of Its Parts? Modeling the Contributions of Language Comprehension Skills to Reading Comprehension in the Upper Elementary Grades. *Sci. Stud. Read.* **2016**, *20*, 436–454. [CrossRef]
33. Cho, S.J.; Goodwin, A.P. Modeling Learning in Doubly Multilevel Binary Longitudinal Data Using Generalized Linear Mixed Models: An Application to Measuring and Explaining Word Learning. *Psychometrika* **2017**, *82*, 846–870. [CrossRef] [PubMed]
34. Cho, S.J.; Gilbert, J.K.; Goodwin, A.P. Explanatory Multidimensional Multilevel Random Item Response Model: An Application to Simultaneous Investigation of Word and Person Contributions to Multidimensional Lexical Quality. *Psychometrika* **2013**, *78*, 830–855. [CrossRef] [PubMed]
35. Goodwin, A.P.; Gilbert, J.K.; Cho, S.J.; Kearns, D.M. Probing Lexical Representations: Simultaneous Modeling of Word and Person Contributions to Multidimensional Lexical Representations. *J. Educ. Psychol.* **2014**, *106*, 448–468. [CrossRef]
36. De Boeck, P.; Wilson, M. A Framework for Item Response Models. In *Explanatory Item Response Models*; Springer: New York, NY, USA, 2004; pp. 3–41.