Author(s): Leontjev, Dmitri; Huhta, Ari; Tolvanen, Asko

Title: L2 English vocabulary breadth and knowledge of derivational morphology : One or two constructs?

Year: 2023

Version: Published version

Article

LANGUAGE
TESTING

Language Testing
1–25

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02655322221114015
journals.sagepub.com/home/ltj
SAGE

L2 English vocabulary breadth and knowledge of derivational morphology: One or two constructs?

Dmitri Leontjev
University of Jyvaskyla, Finland

Ari Huhta
University of Jyvaskyla, Finland

Asko Tolvanen
University of Jyvaskyla, Finland

Abstract

Derivational morphology (DM) and how it can be assessed have been investigated relatively rarely in language learning and testing research. The goal of this study is to add to the understanding of the nature of DM knowledge, exploring whether and how it is separable from vocabulary breadth. Eight L2 (second or foreign language) English DM knowledge measures and three measures of the size of the English vocabulary were administered to 120 learners. We conducted two confirmatory factor analyses, one with one underlying factor and the other treating vocabulary breadth and DM as separate. As neither model had a satisfactory fit without introducing a residual covariance to the two-factor model, we conducted an exploratory factor analysis, which suggested two separate DM factors in addition to vocabulary breadth. Regardless, the analysis demonstrated that the DM knowledge was separate from learners' vocabulary breadth. However, learners' vocabulary breadth factor still explained a substantial amount of variance in learners' performance on DM measures. We discuss theoretical implications and implications for L2 assessment.

Keywords

Constructs, derivational morphology, English as a foreign language, factor analysis, vocabulary

Corresponding author:
Dmitri Leontjev, Department of Language and Communication Studies, University of Jyväskylä, P.O. Box 35, FI-40014 Jyväskylä, Finland.
Email: dmitri.leontjev@jyu.fi

Conceptualising and operationalising vocabulary knowledge and derivational morphology knowledge

Knowledge of derivational morphology (DM), also referred to as morphological awareness or derivational knowledge depending on whether inflectional affixes are involved, has been investigated relatively rarely in a second or foreign language (L2) learning and assessment research. Irrespective of the focus, in our interpretation, these terms imply *analysing words into their constituent parts.* Hence, in our definition, the DM knowledge construct involves *analysing words into stems and derivational affixes and knowing something about the form, meaning, and/or usage of these affixes* (e.g., *build—builder* [somebody who builds]; see Friedline, 2011).

Despite some recent developments in conceptualising and measuring DM knowledge (González-Fernández & Schmitt, 2020; Mizumoto et al., 2019; Sasao & Webb, 2017; Spencer et al., 2015), how it can best be understood and tested is still unclear, including based on neurolinguistic research (Leminen et al., 2019). Influential first language (L1) psycholinguistics research (e.g., Taft, 1981) suggests that L1 English words are analysed into morphemes during lexical access (see also Lieber, 2017). Similarly, Dawson et al. (2021) have recently found evidence for, averaged across age and reading, faster priming not only for morphologically related words but also for pseudo-morphologically related ones (e.g., corn-corner) in L1 English children and adolescents. Furthermore, as Lieber (2017) maintained, both analysing words into morphemes and accessing and processing them as wholes can happen at the same time. This research creates a strong argument for studying DM knowledge and its development. However, the exact relationship between DM knowledge and other kinds of vocabulary knowledge is not clear either. Because of this, how research findings regarding knowledge of derivational morphology can be applied to teaching and assessing word derivation is poorly understood (Friedline, 2011). Adding to the understanding of the construct of DM knowledge is ever so important because it has been found that L2 English derivation poses problems to learners (e.g., Friedline, 2011; McLean, 2017; Schmitt & Meara, 1997), including learners producing inaccurate forms, such as "We have one different" (Friedline, 2011, p. 37). The same appears to be true for young L1 (first language) English learners (Tyler & Nagy, 1989). Thus, the acquisitional processes and stages in relation to DM, including grammatically inaccurate stages or phases of production (sometimes also including overgeneralisations), may be a natural part of language learning; yet, they are poorly understood within L1 and L2 contexts.

To better conceptualise DM knowledge within an L2 context, it seems logical to first conceptualise L2 vocabulary knowledge, of which DM knowledge is a part. Many models of vocabulary knowledge are dimensional. The roots of the dimensional vocabulary knowledge models lie in Richards' (1976) classification of the dimensions of L2 lexical competence into knowledge of the form of the word, its associations, and syntactic properties as well as constraints of use (e.g., Seidenberg & Gonnerman, 2000). Building on Richards, Nation (2001) divided L2 vocabulary knowledge into form, meaning, and use. Word forms can be further divided into spoken and written forms and parts of words; all of these can be known both receptively and productively. In contrast to vocabulary knowledge at a more general level, no accepted dimensional model of DM knowledge exists, at least not a universal model (see, for example, Leminen et al., 2019, for recent psycholinguistic L1 research). However, a small number of studies, elaborated in the following sections, explored whether various measures of DM knowledge and morphological awareness in general load to one or several factors.

Theoretical discussions suggest that DM knowledge is conceptually distinct from other types of word knowledge. Milton and Fitzpatrick (2013), for example, considered knowledge of L2 DM as a part of vocabulary depth (alongside associations), and hence conceptually distinct from at least the breadth of vocabulary knowledge, by which we mean "*knowledge of word forms and primary meaning*" (Koizumi & In'nami, 2020, p. 2). Nation (2001), too, discussed knowledge of L2 word parts as a complex multidimensional construct, itself being an aspect of L2 vocabulary knowledge. Indeed, L2 DM knowledge requires the ability to analyse words into their parts, roots, and derivational affixes, recognising their meanings and whether they change the word class, as well as the ability to use this knowledge in speaking and writing (see Mäntylä & Huhta, 2013; Milton & Fitzpatrick, 2013; Nation, 2001). This suggests that this construct is multidimensional and is a part of the vocabulary depth construct—how well words are known (Henriksen, 1999). Indeed, based on the previous research, dimensions such as morphophonology or morpho-orthography, syntactic knowledge, and semantic knowledge can be parts of the construct of L2 derivational morphology knowledge as well (e.g., Chuenjundaeng, 2006; Friedline, 2011; Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002).

Learners' L2 English DM knowledge has been operationalised, for example, as syntactic function of suffixes with a gap-filling task in the context of separate sentences (e.g., Schmitt & Zimmerman, 2002), with a decontextualised word segmentation task (e.g., Hayashi & Murphy, 2011) or receptive knowledge of meanings of prefixes with a decontextualised non-word task (e.g., Mochizuki & Aizawa, 2000). Some measures, therefore, seem to require learners to focus on derivational morphology more when completing the task, while others elicit learners' breadth of vocabulary knowledge (e.g., their basic meanings) included in the DM task, which might help them substantially to complete the task.

Indeed, research investigating correlations between vocabulary breadth and various DM measures, for example, showed that the magnitude of correlations varies. Hayashi and Murphy (2011) found that the L2 English vocabulary breadth of L1 Japanese learners of English correlated strongly with their performance on an affix elicitation task ($r = .832$ for productive and $r = .842$ for receptive vocabulary) but not with that on a word segmentation task. Schmitt and Meara (1997), however, found only small to moderate correlations ($.27 \leqslant r \leqslant .41$) between L2 learners' receptive vocabulary breadth and their performance on tasks measuring their receptive (learners had to mark all the suffixes that the given words could take) and productive (learners had to write the affixes that could be added to the given words) English DM knowledge.

Friedline (2011) noted that some L2 vocabulary items are processed as unanalysed wholes, whereas in others, derivational affixes are analysed, which might explain differences in the magnitudes of correlations among the studies. In fact, Nation (2001) provided a list of factors that impact the likelihood of learners recognising and using affixes or processing words as wholes, including frequency and productivity of affixes, regularity of use, and semantic transparency.

Regarding studies comparing L1 speakers and L2 learners, Iwaizumi and Webb (2022), for example, compared and contrasted L1 ($n=23$; university students) and L2 English ($n=107$; master, undergraduate, and high-school students) speakers' performance on the decontextualised derivative recall test (Schmitt & Meara, 1997). They found that the performance of L2 learners with larger vocabulary size (3000–5000 headwords) was not significantly different from the performance of L1 speakers, with some L2 learners outperforming L1 speakers. The ability of L2 speakers to recall derived words also varied across groups who acquired different frequency vocabulary. Iwaizumi and Webb (2021) further found that L1 English speakers produce significantly more derived words than English as a foreign language (EFL) and English as a second language (ESL) learners do, and that ESL learners produce more derived forms than EFL learners. There was also an interaction effect with vocabulary frequency bands, meaning the higher the vocabulary size was, the more derived words the learners produced, even if the magnitude varied among the three groups (L1, ESL, and EFL).

Our hypothesis is that the measures used in previous studies elicited both L2 vocabulary breadth (and other aspects of vocabulary knowledge) and L2 DM knowledge. The interaction between these in the measures may have led to different, even conflicting, findings (see also Leontjev et al., 2016). At a more general level, the different correlations across studies raise the question of whether vocabulary breadth and DM knowledge are separate constructs.

We next review studies that have specifically investigated the dimensionality of vocabulary knowledge, arguing for either a unified construct or several separate constructs. We note that there are not many such studies to date.

Literature review

Studies implying the inseparability of DM knowledge and vocabulary breadth

Two fairly recent empirical studies employing Structural Equation Modelling (SEM) suggested that DM knowledge is a part of the same construct as other vocabulary measures, such as vocabulary breadth and depth. In the L1 English context, Spencer et al. (2015) used confirmatory factor analysis (CFA) with several vocabulary depth and breadth measures as well as a set of measures of DM knowledge with L1 speakers of English in two studies, one with 99 fourth-grade students (aged 9–12) in the United States and a follow-up study with 90 eighth-grade students (no age given) in the United States. There were nine morphological awareness tasks in the first study, including five tasks eliciting DM knowledge (involving real words, non-words, and improbable suffixes), three compounding tasks, and one inflectional morphology task. The vocabulary measures were the vocabulary subtest from the Stanford-Binet (participants defined orally words presented to them) and the Peabody Picture Vocabulary Test (participants pointed at the pictures of the words presented orally). (See Spencer et al.'s paper to learn more about their measures.) Both measures were, therefore, evaluating vocabulary breadth. In the second study, the same 23 words were used in vocabulary breadth (e.g., asking the participants what the words meant), depth (e.g., asking the participants to provide words that meant the same), and morphological awareness tasks, including modifying the source words to complete sentences requiring the use of inflectional or derivational affixes and forming new words from source words. In both studies, the authors found that the measures loaded into a single factor. This led the authors to suggest that DM knowledge is an inseparable part of vocabulary knowledge. However, we note, as the authors themselves suggested, that Spencer et al.'s (2015) study had a somewhat small sample size, which could have resulted in the failure to find a significant difference between the one- and the two-factor model. In addition, we wonder whether the analysis without the improbable suffixes task, considering its low correlations with the rest of the measures, could have resulted in a different outcome regarding the one- versus two-factor model.

In the L2 context, González-Fernández and Schmitt (2020) conducted an SEM study with 144 Spanish-speaking learners of English to explore the word knowledge construct(s) using measures of breadth and depth of vocabulary. The measures included form recall and meaning recognition in the form-meaning link component, form recall and form recognition of derivatives, meaning recall and recognition of multiple meanings, and form recall and recognition of collocates. The authors used the same 20 words across the measures. They found that the one-factor model with residual covariances introduced between recognition and recall for each pair of measures eliciting the same component had the best fit, suggesting that there is one construct of vocabulary knowledge including DM knowledge. However, González-Fernández and Schmitt's (2020) model had only two indicator variables for the knowledge of derivatives (and in the other three components), which could have affected their results (see Raubenheimer, 2004).

Studies implying the separability of DM knowledge and vocabulary breadth

Differently from the studies outlined above, there are studies that imply that DM knowledge and vocabulary breadth are separate constructs. We outline two such recent studies, one exploring the separability of vocabulary depth (to which DM knowledge belongs) from breadth and a series of studies exploring the dimensionality of morphological knowledge.

With regard to L1 English, Goodwin and her colleagues investigated the dimensionality of morphological knowledge and its relationship with reading quite extensively in the United States (e.g., Goodwin et al., 2012, 2017, 2021; Goodwin, Petscher, & Tock, 2020; Goodwin, Petscher, Tock, McFadden, et al., 2020). Their studies included a range of measures of vocabulary depth and morphological awareness (e.g., antonym/synonym, word relations and polysemy measures, as well as suffix choice in real and pseudowords and morphological judgement tasks) but also some measures tapping vocabulary breadth (e.g., Gates-MacGinitie Standardized Test of Reading Vocabulary and Woodcock Language Proficiency Battery Picture Vocabulary), although they did not use these terms when describing their instruments (nor did they refer to L2 vocabulary research). Most of their vocabulary depth measures clearly elicit DM knowledge. Here, we review their two most recent studies. In the first study, Goodwin et al. (2017) investigated 371 seventh and eighth graders who completed seven written morphological measures and a measure of vocabulary knowledge and reading comprehension. The morphological measures included suffix choice tasks based on either real or pseudowords, tasks requiring decisions about the morphological similarity of pairs of words, tasks involving reading aloud and spelling of derived and root words, as well as self-assessing the meaning of derived and root words. The dimensionality of morphological knowledge was investigated via CFA. The researchers concluded that morphological knowledge can be conceptualised as a general construct and several specific dimensions, each uniquely related to literacy skills. Recently, Goodwin et al. (2021) investigated 3214 students from grades 5 through 8 (8% of which were English language learners) who completed a variety of morphological tasks, including tasks requiring comparison of words that included or lacked shared morphemes, filling out gaps in sentences with an appropriate word form, identifying the meaning of morphologically complex words in sentences, spelling heard words, and identifying the correct pronunciation of seen words. The researchers used multiple-group item response modelling and CFAs to study the dimensionality of morphological knowledge. They found that it consists of four empirically separable skills: Morphological Awareness, Morphological-Syntactic Knowledge, Morphological-Semantic Knowledge, and Morphological-Orthographic / Phonological Knowledge.

The research by Goodwin and colleagues suggested, then, that morphological knowledge itself may not be a single construct but may consist of several constructs. Since most of their tasks clearly measured knowledge of derivational morphology, their results suggest that English vocabulary knowledge is not unidimensional.

In the L2 context, Koizumi and In'nami (2020), in an SEM study of 255 Japanese learners of L2 English aged 18 and above, investigated whether vocabulary depth is separate from breadth. The study included one measure of vocabulary breadth based on the JACET8000 vocabulary list, which the authors divided into three parts based on frequency bands, and four measures of vocabulary depth: a word association measure, two polysemy measures, and a collocation measure. The authors found that vocabulary breadth and depth (to which DM knowledge is considered to belong) were two separate factors although these factors correlated strongly both in conventional SEM ($r=.945$) and in Bayesian SEM ($r=.943$). One limitation of the study was, as the authors noted, that the sample was limited to Japanese learners of English. Furthermore, their measures were limited in that the questions were in the multiple-choice format.

To summarise, it appears empirical research into the dimensionality of learners' vocabulary knowledge is inconclusive. Our study adds to this line of research by investigating whether L2 vocabulary breadth and derivational morphology knowledge are separable even if correlated. We examine this by using SEM as the analytical framework, namely, CFA (see In'nami & Koizumi, 2011; Tarka, 2018, on using SEM in applied linguistics). Then, if the analyses indicate that our data support the separability of DM knowledge and vocabulary breadth, we also investigate the relative contribution of learners' L2 English vocabulary breadth and DM knowledge in these measures.

Methodology

Research questions

The research questions we addressed were the following:

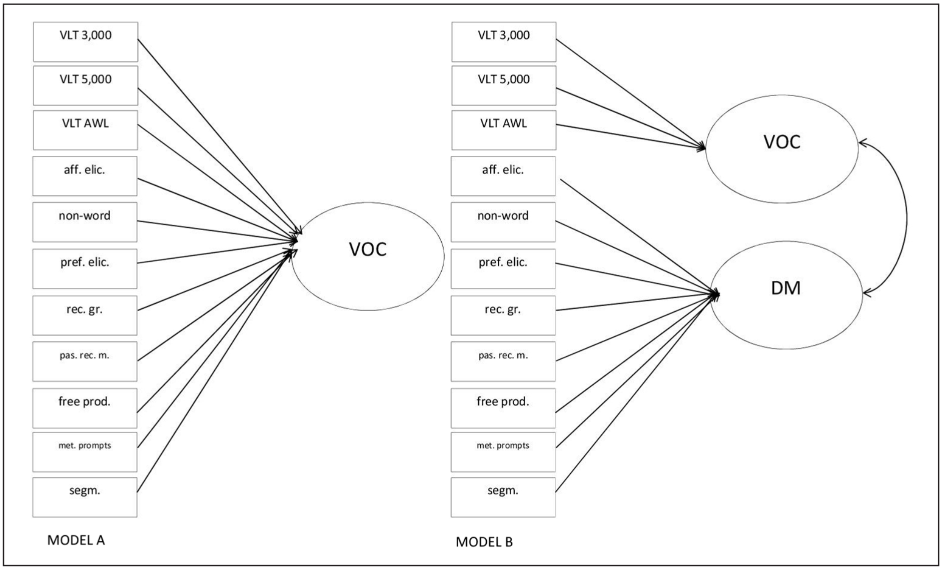

Figure 1. Hypothesised latent constructs behind the measures.

1. *Research Question 1.* Can applied linguists and language assessors empirically separate the construct of L2 English derivational morphology knowledge from vocabulary breadth in measures eliciting the former, or are they inseparable?
2. *Research Question 2*. If there are two or more constructs underlying the measures, to what extent do different measures that have been used to tap L2 English derivational morphology also elicit vocabulary breadth?

Based on the previous research, we hypothesised that two alternatives are likely: (1) all the measures in the study load onto one construct or (2) measures eliciting vocabulary breadth should form a construct reliably separable from measures tapping into learners' derivational morphology knowledge. Previous research suggests that two hypothesised models are likely (Figure 1): a one-factor model (model A) and a two-factor model in which the two factors correlate (model B).

Participants and data

The participants were 10 different intact groups totalling 120 learners, in year 10 (the first year of senior secondary education) in schools in Finland (n=71) and Estonia (n=49), with a median age of 17, and the age range being 15–18 years.

We deemed the sample as sufficient for both CFA and exploratory factor analysis (EFA). The minimum recommended sample for both is 100 and the alternative rule of thumb of 10 cases per variable (Plonsky & Gonulal, 2015; Wolf et al., 2013). Furthermore, MacCallum et al. (1999) suggested that a sample of at least 100 is sufficient for factor analysis with 3–4 indicators per factor, communalities of about .5, and factor loadings of

.8 or above. Finally, in addition to these rules of thumb, we considered the outcomes of the simulation study by Wolf et al. (2013), which suggested a sample size of 120 for two factors with loadings of .65 and 6 indicator variables and the same sample for loadings of .8 and 3–4 indicators (see the "Results" section). We also considered factor determinacies when making the decisions.

We offered feedback to the learners and teachers (with the learners' permission) about the learners' performances as an incentive to participate in the study. The participants were selected from the level of education that would suggest they were at least at the B1 level in English. This was based on Nation's (2001) argument that learners are ready to be taught derivational affixes at the lower-intermediate level of L2 proficiency, which corresponds to the B1 level on the CEFR scale (Common European Framework of Reference; Council of Europe, 2001). On average, the participants fulfilled this sampling criterion −85.5% of the sample rated at level B1 or above based on Multifaceted Rasch analysis of two writing samples per learner, each rated independently by at least two raters on the CEFR scale. There was no significant difference in the learners' English proficiency between the two countries, nor were there any in the learners' performance on any of the measures in the study when a Bonferroni correction was applied to account for the family-wise error. Without the correction, the *t*-tests indicated a significant difference in the Vocabulary Levels Test (VLT) 5000 and the non-word affix elicitation and word segmentation tasks. These differences were small, for example, for the VLT 5000 ($d=0.39$), except for the non-word affix elicitation ($d=0.53$) (see the following section for the description of the measures). As our intention was not to compare groups of learners but rather to study the construct with learners of roughly similar proficiency, and because the big picture was that the groups in the two countries were very similar in their performances, we considered them together.

Measures

Vocabulary breadth measures. The learners completed three levels (3000 words, 5000 words, and Academic Word List [AWL]) of the Finnish, Estonian, and Russian bilingual adaptations of Schmitt et al.'s (2001) VLT created by the project team to which we belonged. Namely, instead of providing definitions in English for the items, we provided Finnish, Russian, or Estonian equivalents or definitions. For example, for the first item of the VLT 3000 against the six options in English, the Finnish learners saw the three words "ajatus" "kämmen," "vyö," the first of which is the direct translation of "idea" (the word used in the original) into Finnish, and the other two, the equivalents of "palm" and "belt" in Finnish. We thus followed the approach used in other versions of bilingual levels tests.

We used breadth measures to elicit learners' performance that did not require the learners to analyse words into their constituent parts, affixes, and root morphemes (while indeed some of the VLT items contain affixes, for example, "fragrant" or "professional," DM knowledge is not elicited in the VLT; see Supplementary Materials). Furthermore, these measures have frequently been used to represent the breadth dimension of vocabulary knowledge (Hayashi & Murphy, 2011; Mochizuki & Aizawa, 2000). Bilingual versions of the VLT were designed for several reasons, which are detailed in the

Table 1. Descriptive statistics.

Measure	*N*	Mean	*SD*	Min/Max	Cronbach's alpha
VLT 3000	120	22.12	6.03	6/30	.89
VLT 5000	120	17.84	5.78	4/30	.85
VLT AWL	120	19.86	5.98	4/30	.88
Affix elicitation	116	8.95	4.06	0/15	.87
Non-word affix elicitation	116	4.02	3.49	0/11	.85
Prefix elicitation	117	6.62	2.96	0/12	.79
Grammar recognition	116	6.44	2.38	1/10	.73
Passive recognition of the meaning	114	6.29	2.23	1/10	.61
Free production, no. of der. Affixes	119	7.66	4.55	1/29	.80
Metalinguistic prompts, no. of der. Affixes	116	5.39	4.48	0/18	.76
Word segmentation, no. of der. Affixes	108	13.31	5.46	0/29	.80

Note: VLT: Vocabulary Levels Test; AWL: Academic Word List.

Supplementary Materials. To summarise, the bilingual versions made the administration of the VLTs more practical and may have increased their validity as the learners could use their L1 to indicate their knowledge of the English words targeted by the tests.

We analysed the bilingual VLTs together with the rest of the measures using Winsteps (Linacre, 2015). Such Rasch analyses were also among the validation approaches used by Schmitt et al. (2001). The person separation reliability for all the three VLT scales showed that they reliably separated low-performers from high-performers. The lowest person reliability was observed with the 3000 band of the VLT, .79 with extreme cases. All three scales taken together could reliability separate the sample into 3 to 4 groups, with the person reliability with extreme cases being .93. The 3000 VLT was the easiest and the 5000 VLT the hardest, with the AWL coming in between them, which corresponded to Schmitt et al. (2001) results.[1] See Table 1 for descriptive statistics and Cronbach's alphas. See Supplementary Materials for the separation reliability and strata for persons and items.

Measures of L2 English DM knowledge. The DM measures included those eliciting learners' productive and receptive DM knowledge (see e.g., Carlisle, 2000; González-Fernández & Schmitt, 2020, for rationales). The tasks also either measured the learners' DM knowledge in context or as separate words. Finally, the tasks measured different aspects of the learners' DM knowledge, for example, the semantics of affixes or syntactic knowledge of affixes. The source and target words in the measures spanned the same VLT levels or below those selected as the vocabulary breadth measures in the study. The source words were generally within the first 3000 bands of the VLT; in single cases, the source words were from the 4000 to 7000-word bands. This was done to ensure that the learners could recognise the base words in the tasks (with the exception of, naturally, the non-word affix elicitation task; see below).

The derivational affixes spanned Bauer and Nations' (1993) affix levels 3 to 7, with the exception of the metalinguistic prompts task and the free production task, where we could not control the affixes the learners used. For the most part, however, the affixes were from levels 3 to 5 (i.e., frequent and/or regular derivational affixes). Only six words were used several times across the measures. The learners first completed the VLT measures, followed by the decontextualised measures, followed by the rest of the measures. We next give a brief overview of the measures. The measures in full have been given in the Supplementary Materials to this paper.

- **Affix elicitation task** (*k*=15; example item: "I am sure the company will hire him. He will . . . (varmasti) get a summer job"); recalling frequent derived words, completing sentences using an affix, contextual clues (see Friedline, 2011; González-Fernández & Schmitt, 2020; Hayashi & Murphy, 2011; Schmitt & Zimmerman, 2002), and L1 translation.
- **Non-word affix elicitation task** (*k*=13; example item: "She could bourble animals very well because she was a good . . . bourble . . . .") was similar to the affix elicitation task, but non-words were used instead, and an explanation in L1 was provided (e.g., "an action that the highlighted word describes"; see Mäntylä & Huhta, 2013).
- **Prefix elicitation task** (*k*=12; example item: "She never says directly what she means. She is always very . . . direct.," 22 prefixes provided for the learners to select from); recognising the appropriate prefix (focusing on semantics of prefixes) for base words in sentence contexts; base words belonged to the first 2000 words on the VLT lists as well as the AWL (see Mäntylä & Huhta, 2013).
- **Grammar recognition task** (*k*=10; example item: "The men . . . the road. Multiple-choice options: widen; wideness; widely"); recognising the derived form that fits the sentence grammatically (in terms of parts of speech) (see Akande, 2003; Mochizuki & Aizawa, 2000; Spencer et al., 2015).
- **Recognition of the meaning task** (*k*=10; example item: "He seems to be faultless. Options: perfect; full of problems; usual"); recognising the meaning of the derived word; the format is somewhat similar to Mochizuki and Aizawa (2000) who used single non-words instead of derived words in sentence context, and closely resembles González-Fernández and Schmitt's (2020) task format.
- **Free production task** (*k*=10; example item: AGREE); producing derived words based on base words given to learners as a decontextualised list; derivational affixes in the correctly formed derived words were counted, for example, *agreement*, *agreeable* resulting in the score of two (Schmitt & Meara, 1997).
- **Metalinguistic prompts task**, (*k*=10; example item: "HORROR: Noun [e.g., farmer] . . . Verb [e.g., go]: . . . Adjective [e.g., good] . . ."); producing one noun, one verb, and one adjective from the given word; derivational affixes were counted in the correctly formed words belonging to the correct part of speech (see Schmitt, 1998, who, however, used the oral modality)
- **Word segmentation task** (*k*=49; excerpt: "It was, however, an extremely difficult 'make-up', if I may use such a theatrical expression . . ."); recognising derived words in a coherent text and marking the affixes in these words. This task was completed on paper, as we found that completing the measure on the computer was difficult for the learners.

We reviewed and trialled the measures in several stages, including a pilot study with 22 university learners of English, whose proficiency was estimated at about level B2 on the CEFR scale. After piloting, we made adjustments to the wordings and instructions. We also conducted classical and modern (Item Response Theory; IRT) item analysis of the measures, both during the piloting and with the dataset in this study. The measures worked reasonably well, except for one measure we excluded from this study, as the IRT item analysis indicated that its reliability was low, and it had several outfitting items. The classical item analysis indicated that all the items in the measures selected for study worked well (had respectable facility and discrimination indices) except for one item in the grammar recognition task, which had a facility value of .91, and one item in the non-word affix elicitation that had a facility value of .16 (which, however, discriminated reasonably well at .33). The items in the measures selected for the study had facility values ranging from .22 to .87 and discrimination (based on 27% top and 27% bottom) ranging from .27 to .96. Upon consideration, we decided not to exclude the two items with high and low facility from the scales.

Procedure

The learners completed most of the tasks in a Web-based assessment system. The data collection took place under controlled conditions. The learners completed all of the tasks save for the word segmentation task in one session. The word segmentation task was completed separately in most groups, within a week, due to its different modality, the time constraints, and to avoid fatigue. In some groups, researchers were present during the data collection, monitoring the learners' performance together with teachers. When the researchers could not monitor the data collection, detailed instructions to the teachers, including prevention of cheating and addressing the learners' queries, were written. While the data collection happened mainly during the regular school lessons, ample time—one double period, that is, 1.5 hours and a break—was given to the learners to complete the tasks, and all learners completed the tasks within that time. No sign of learner fatigue was observed, although a few learners seemed to show unusual behaviour (e.g., lack of engagement or suspected cheating) when working on the tasks as indicated in our observation notes. We checked these responses as well as calculated Mahalanobis distances across all the measures. Only the performances of the learners (only a few cases) which emerged as multivariate outliers *and* exhibited unusual behaviours according to observation notes were removed from the dataset.

Analysis

The learners' performances on the productive measures were scored such that minor misspellings in the stem of words that the learners made were not penalised (misspellings in affixes were). That is, we accepted such responses as **unbelievable* as correct responses, as the learners correctly supplied affixes in these cases and spelled these correctly. However, we considered cases as *unbelieve* in place of *unbelievable* as incorrect.

The analyses used to respond to the research questions were conducted with Mplus version 7.3 (Muthén & Muthén, 1998–2015) with robust multiple likelihood as the estimator. As some data were missing, we used the full available information for calculating the estimates. We followed Hu and Bentler’s (1999) criteria for good model fit, namely, comparative fit index (CFI) and Tucker–Lewis index (TLI) higher than .95, root mean square error of approximation (RMSEA) smaller than .06 (smaller than .08 still acceptable), and standardised root mean square residual (SRMR) smaller than .08. The chi-square test should, too, be preferably non significant. We also calculated factor score determinacy values to study the validity of models. We ran two CFA models that were possible based on previous research (see Figure 1; the first research question). We followed the analysis by estimating the amount of variance in the measures explained by vocabulary breadth (the second research question). However, as we will elaborate below, we also ran an EFA in the Mplus environment using robust multiple likelihood as the estimator as an add-on, exploratory component to this study. We will report our rationale for the number of extracted factors, the rotation method, the factor loading matrix, and other data useful to confirm the constructs that DM measures tap into. In addition, we will report the fit indices and compare the differences between models statistically. We built mainly on the comprehensive review by Plonsky and Gonulal (2015) as our rationale for (a) making decisions about analyses preceding the EFA proper (e.g., whether data are factor analysable), (b) selecting the number of factors to extract, and (c) reporting and interpreting data. We first studied the correlations matrix for patterns, followed by calculating by hand and then studying the Kaiser–Meyer–Olkin (KMO) statistic, because it is not obtainable in Mplus in SPSS, and then by studying the item-to-participant ratio. We decided on the number of factors to extract based on the scree plot (Figure 3) and the chi-square test comparing the three-factor model fit to that of the two-factor model rather than the Kaiser-1 rule (Eigenvalues >1) rule of thumb, which tends to underestimate or overestimate the number of factors (Plonsky & Gonulal, 2015; Russell, 2002). We also studied the communalities, not directly obtainable in Mplus but possible to calculate as 1—residual variance, when considering whether the measures should be kept in the EFA. Considering the theoretical understanding of the constructs, we used an oblique rotation method in Mplus—geomin—to allow the extracted factors to correlate. We finally studied the pattern matrix and the structure matrix to interpret the solution.

To answer the second research question, based on the second CFA model, we calculated the variance explained by the vocabulary breadth factor. In the post hoc EFA analysis, we studied the factor structure, that is, the correlations of the measures with the factors.

The following section will first focus on the two CFA models, studying the two-factor model more closely. In the second subsection, we propose an alternative insight into the constructs underlying the data focusing on the EFA.

Results

The descriptive statistics and the internal consistency of the measures are presented in Table 1. The internal consistency of the two multiple-choice tasks was somewhat low, though acceptable. We also checked for multivariate outliers in the data and found that the Mahalanobis distances were below the critical value for Chi-square of $\chi(11)=31.264$, $\alpha=.001$ (see Kline, 2011) and found that there were no multivariate outliers. To give a better overview of the data, we provide correlations between measures in Table 2.

Table 2. Pearson's correlations among raw measures.

Measure		VLT 5000	VLT AWL	Aff. elic	Non-word aff. elic.	Pref. elic.	Gr. rec	Pass. rec. mean	Free prod.	Metalng prompt	Segm.
VLT 3000	Corr.	.795	.848	.586	.530	.578	.652	.526	.435	.538	.089
	Sig.*										.369
VLT 5000	Corr.		.822	.633	.558	.607	.624	.550	.467	.524	.184
	Sig.*										.049
VLT AWL	Corr.			.600	.568	.577	.606	.546	.474	.534	.174
	Sig.*										.066
Affix elicitation	Corr.				.554	.618	.629	.590	.536	.488	.217
	Sig.*										.024
Non-word aff. elicit.	Corr.					.616	.609	.709	.556	.599	.373
	Sig.*										
Prefix elicitation	Corr.						.657	.665	.550	.467	.111
	Sig.*										.272
Grammar recognition	Corr.							.647	.576	.548	.212
	Sig.*										.025
Pass. rec. meaning	Corr.								.473	.474	.184
	Sig.*										.054
Free production	Corr.									.674	.327
	Sig.*										
Metaling. prompts	Corr.										.354
	Sig.*										

Note: VLT: Vocabulary Levels Test; AWL: Academic Word List.

*Unless otherwise stated, $p < .001$.

Table 3. Fit indices of the CFA models.

	χ^2	df	Sig.	CFI	TLI	RMSEA	SRMR
One-factor	176.06	44	<.001	.843	.803	.158	.07
Two-factor	75.84	43	.002	.961	.950	.08	.052
Two-factor with res. covariance	58.88	42	.044	.980	.974	.058	.047

Note: CFI: comparative fit index; TLI: Tucker–Lewis index; RMSEA: root mean square error of approximation; SRMR: standardised root mean square residual; CFA: confirmatory factor analysis.

It emerges from Table 2 that with the exception of the word segmentation task, all the measures intercorrelated moderately to strongly, but there were stronger intercorrelations among the three vocabulary breadth measures. The word segmentation task correlated with most of the measures weakly (with many correlations being non-significant) but correlated moderately with the free production and the metalinguistic prompts tasks. The correlation between the free production and the metalinguistic prompts task was also the largest compared to their correlations with other measures. Therefore, the correlational matrix suggested that there were some clear patterns with regard to the shared variance in the data.

The following results will be presented in two subsections, one focusing on the CFA and the other, accounting for the lack of fit in the CFA models, on the EFA.

Testing theoretically informed models: CFA

The fit of the one-factor model in the CFA analysis was very low (see Table 3). Furthermore, the modification indices suggested substantial residual covariances among all of the vocabulary breadth measures. In other words, in the one-factor model, it emerged that there was a substantial shared variance across all of the vocabulary breadth measures that the single factor did not account for. The obvious interpretation of this finding was that a separate latent construct was behind the vocabulary breadth measures. That is, while introducing these large residual covariances would have improved the fit substantially, this would have also violated the parsimony principle (explaining the data with the least number of parameters) and, in fact, would mean we ignored the obvious interpretation of these residual covariances as an invitation to study the two-factor model instead.

The two-factor model had a much better fit than the one-factor model. However, it was still only marginally acceptable, judging by a combination of the fit indices. As the first step, we studied the suggested modification indices. Mplus suggested only one modification index: a residual covariance between the free production and the metalinguistic prompts task. Our explanation for the two measures having some variance not accounted for by the model (the DM factor they loaded onto) is that, unlike the rest of the DM measures in the study, these two measures elicited decontextualised knowledge of DM. Considering that there is a theoretically informed explanation for the residual covariance and that this minimum modification substantially improved the fit, providing balance between parsimony and fit, we added it to the model. The model with two factors (the

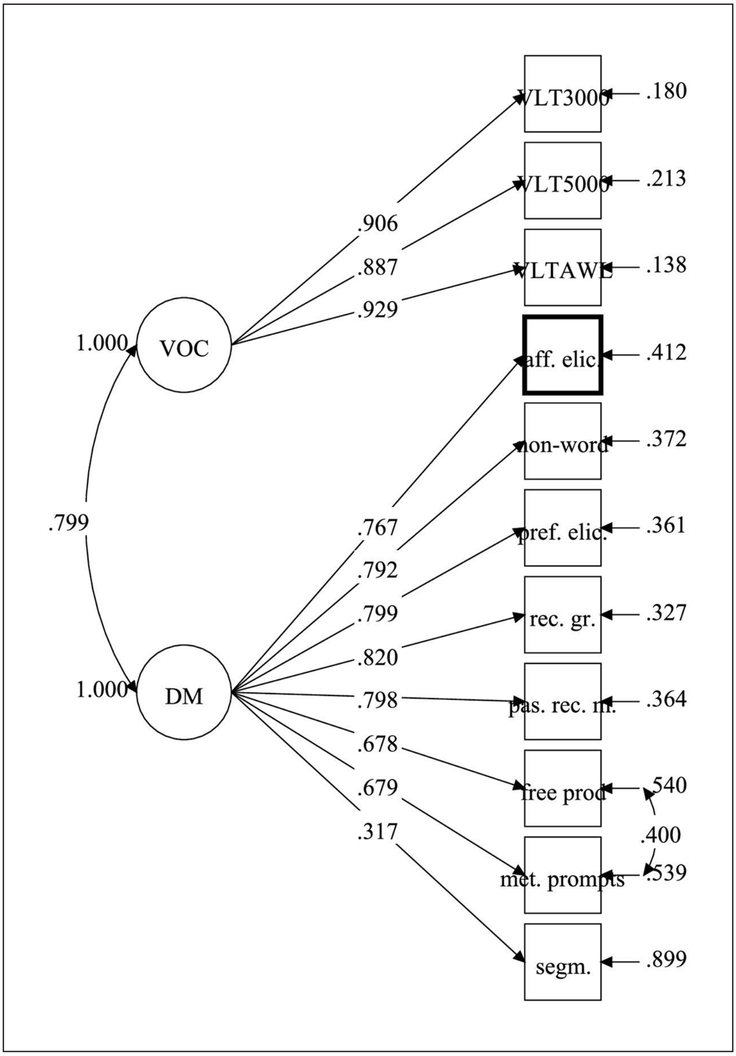

Figure 2. Confirmatory factor analysis: two-factor model.

vocabulary breadth and the DM knowledge) and the residual covariance resulted in a good fit (Table 3). The standardised loadings for the model are presented in Appendix 1. The following Figure 2 illustrates the model.

It appears, therefore, that there is a high degree of certainty that in our dataset, the DM factor is separable (and separate) from the vocabulary breadth factor, even if correlated with it.

To answer the second research question, we studied how much of the variance in our DM measures was explained by the vocabulary breadth factor (Table 4).

Table 4. Amount of variance in the measures explained by the vocabulary breadth knowledge, two-factor model.

	Vocabulary breadth factor r^2
Affix elicitation	.38
Non-word affix elicitation	.40
Prefix elicitation	.41
Grammar recognition	.43
Passive recognition of the meaning	.41
Free production	.29
Metalinguistic prompts	.29
Word segmentation	.06

It, therefore, emerges from the analysis that, across all DM measures, substantial variance was explained by the vocabulary breadth factor.

Exploratory factor analysis

This section presents a second exploratory way of addressing the marginal fit of the second of the two theoretically informed models as detailed in the previous section; that is, through studying the data new using EFA conducted in the Mplus environment.

As the first step, we studied the KMO and the item-to-participant ratio to check if the data were suitable for an EFA, which suggested that they were: KMO = .915 and items-to-participant ratio, 10.9. As Plonsky and Gonulal (2015) recommended, we followed several criteria in making the decision regarding the number of factors to extract. First, we studied the scree plot, which showed that there are at least two, and possibly even three, factors (see Figure 3).

In addition, we studied and compared the fit of the models. The analysis indicated that a three-factor model fit the data significantly better than the two-factor model, as demonstrated by a chi-square test comparing the three-factor model against the two-factor model, $\chi^2(9) = 21.043$, $p = .0125$. Based on this combination, we extracted three factors. The fit of the model was good, RMSEA = .087; CFI = .973; TLI = .941; SRMR = .025; $\chi^2(25) = 47.45$, $p = .004$.[2] The communalities were also substantial, save for the segmentation task, ranging from .855 to .563, the value for the segmentation task being on the lower side, .235. Still, it was above .2 (Child, 2006), so we kept the item. Factor determinacy values were also strong, .970, .950, and .912, respectively, for the three factors.

The factor pattern matrix is presented in Table 5.

The analysis showed that there were some cross-loadings; for example, the affix elicitation task loaded onto the second factor and additionally onto the first factor, and the non-word affix elicitation task loaded onto the second factor and additionally loaded onto the third factor. However, these were lower than .30. Apart from the affix elicitation task, the loadings onto the corresponding factor were large. However, as the loading of this measure is well above .30, it can be considered significant (Plonsky & Gonulal, 2015). The factors were also, as hypothesised, substantially intercorrelated, with Factor

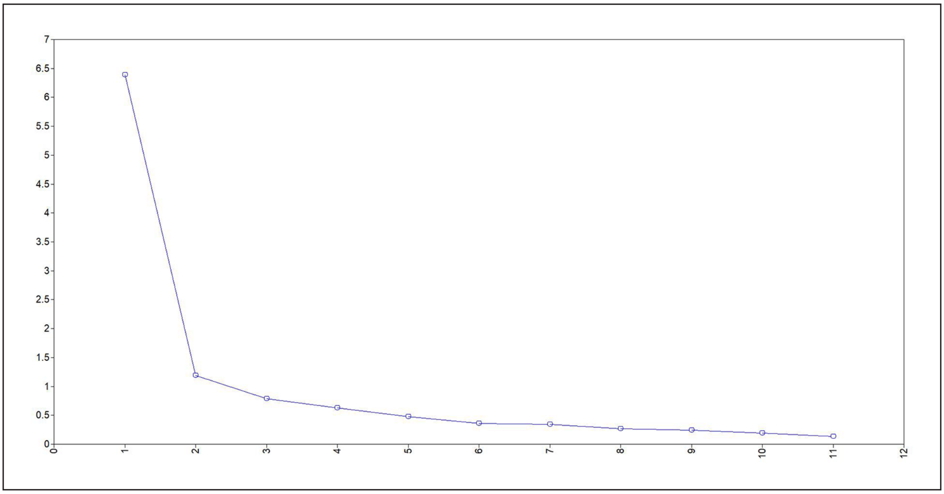

Figure 3. Scree plot.

Table 5. Pattern matrix of the three-factor model.

Measures	Geomin rotated loadings		
VLT 3000	.927	.006	–.028
VLT 5000	.808	.103	.000
VLT AWL	.920	–.002	.010
Affix elicitation	.295	.428	.110
Non-word affix elicitation	.016	.632	.228
Prefix elicitation	.160	.687	–.012
Grammar recognition	.246	.512	.131
Meaning Recognition	–.015	.977	.147
Free production	–.006	.150	.703
Metalinguistic prompts task	.148	–.010	.749
Word segmentation	–.165	.025	.546

Note: VLT: Vocabulary Levels Test; AWL: Academic Word List.

1 correlating at *r*=.700 with Factor 2 and at *r*=.573 with Factor 3, and Factor 2 correlating with Factor 3 at *r*=.695. Notable correlations were also observed between the measures and the factors, as shown in Table 6.

That is, with the exception of the word segmentation task, the measures correlated strongly with all three factors.

Regarding factor interpretation, considering the strong loadings of all the vocabulary breadth measures, Factor 1 was the vocabulary breadth factor. The other two factors have to do with DM knowledge, but it is not entirely clear why there are two factors. However, we will try to interpret them in the following section. Regardless of the interpretation, the

Table 6. Structure matrix for the three-factor model.

Measures			
VLT 3000	.915	.636	.509
VLT 5000	.880	.669	.536
VLT AWL	.925	.650	.537
Affix elicitation	.659	.712	.577
Non-word affix elicitation	.589	.801	.676
Prefix elicitation	.634	.791	.557
Grammar recognition	.680	.775	.628
Meaning Recognition	.585	.865	.524
Free production	.503	.634	.804
Metalinguistic prompts task	.571	.614	.827
Word segmentation	.166	.289	.468

Note: VLT: Vocabulary Levels Test; AWL: Academic Word List.

findings clearly indicate that the vocabulary breadth measures did not load onto the same factor as the DM knowledge measures.

Discussion

Our aims with this study were (a) to find out whether there was a common factor behind different measures of learners' derivational morphology knowledge, which was separate from their vocabulary breadth, and (b) to study how much variance in the measures was accounted for by the DM knowledge factor and how much of vocabulary breadth it explained. Based on previous research, we ran two CFAs, one assuming vocabulary breadth and DM knowledge form one factor, and the other that they form two factors. To account for issues with model fit, we ran an EFA, which can also pave the way for new CFA studies with new datasets.

The analyses demonstrated that the DM measures belonged to a different construct (or constructs) from vocabulary breadth. This adds validity to studies that conceptualise derivational morphology knowledge (or depth of vocabulary) and vocabulary breadth as separate, albeit related constructs (Goodwin et al., 2017; Koizumi & In'nami, 2020). Furthermore, Goodwin's findings about the multidimensional nature of morphological knowledge can explain the emergence of the third factor in the EFA analysis that we conducted.

Informed by Goodwin's (e.g., Goodwin et al., 2017, 2021) research, there is a possibility that a two-order factor, three-factor, or a specific factor model can reflect the construct of the DM knowledge. Further research is needed to shed more light on whether there are indeed more factors to which DM measures load by confirming our EFA model in a CFA study with new data. Alternatively, the third factor in the EFA model might not be confirmed with future data, and a two-factor model may be confirmed instead.

Three of the DM measures (metalinguistic prompts, free production, segmentation) appeared to load onto a different factor from the rest of the DM measures in the

EFA model (see Table 5) in this study. It is not clear to us why these three DM measures may differ from the others purportedly tapping the same construct. We speculate that the extensive textual context of the segmentation task is responsible for it involving different processing and task-taking strategies than the other tasks. For their part, metalinguistic knowledge tasks may require different types of knowledge from the other DM measures. Furthermore, free production is the most open-ended of our tasks and may tap into learners' willingness to persevere and produce more and more words (i.e., derived forms). However, it is difficult to see what could be common to all three of these tasks and what sets them apart from the other DM measures. Finally, it is quite possible that the DM knowledge factor formed by the three measures in the EFA analysis is not real but some artefact of the particular dataset and that all DM measures form one factor.

The EFA provides additional insights regarding the measures. The cross-loading of the affix elicitation on the vocabulary breadth factor is notable, even if below the threshold of .30 (Table 5). We suggest that for this task, the L1 cues given to the learners resulted in that learners were likely to use their knowledge of words in addition to their DM knowledge. We will return to this later in our discussion of the findings pertaining to the second research question. Further research with a larger dataset and using confirmatory factor analysis to validate the proposed three-factor model is needed to see whether this and other cross-loadings would dissipate in such a study.

We should also note that, as hypothesised based on the previous research, our models showed that the DM knowledge factor(s) strongly correlated with the vocabulary breadth factor. We, however, following Koizumi and In'nami (2020), do not consider a strong correlation as implying that there is one factor behind the measures, as also our analyses showed.

Our finding is, indeed, counter to other recent SEM studies (González-Fernández & Schmitt, 2020; Spencer et al., 2015). However, as we reported earlier, both Spencer et al. (2015) and González-Fernández and Schmitt (2020) listed limitations that could have accounted for their particular findings. To summarise, their main limitations were a limited number of measures of DM knowledge in González-Fernández and Schmitt (2020), $k=2$, and a relatively small sample size Spencer et al. (2015), $n=99$ and $n=90$ in study 1 and study 2, respectively. In the present study, more measures were tapping into the DM knowledge construct than in González-Fernández and Schmitt's (2020) study, which, we propose, allowed for us to trace the separability of the two constructs, which González-Fernández and Schmitt (2020) did not. The relatively small sample size in Spencer et al. (2015) may also have contributed to the different results.

Regarding research question two, it emerged that learners' vocabulary breadth was strongly involved in the learners' performances on all our DM measures. In the two-factor CFA model, apart from the word segmentation task, vocabulary breadth accounted for 29% to 43% of the variance in the measures. A similar picture emerged from the EFA analysis (Table 6); that is, the DM measures substantially correlated with the vocabulary breadth factor in the EFA.

These results suggest that DM knowledge research findings should be interpreted taking into account the possibility that learners' vocabulary breadth plays a role in their performance on DM measures (see also McLean, 2017). There is a possibility that even if learners' DM knowledge is separate from other components of word knowledge, such as the breadth of vocabulary, both processing vocabulary items as wholes and analysing them could be involved when learners perform on DM measures, as L1 and L2 research has shown (e.g., Iwaizumi & Webb, 2021; Lieber, 2017). That is, in their performance on tasks designed to elicit DM knowledge, learners can refer to both their DM knowledge *and* vocabulary breadth knowledge.

Our findings have a number of implications for the assessment of L2 English DM knowledge, both in the classroom and in further research. We will elaborate on these in the following sections.

Conclusion

The aim of the present study was twofold (a) to find out whether there was a common construct underlying measures often used to operationalise learners' L2 English knowledge of derivational morphology (DM) which is separate from vocabulary breadth and (b) to determine how much of the variance in the measures used in the study is explained by the vocabulary breadth and the DM knowledge factors.

The results suggested that DM measures loaded onto a single construct, or potentially, several of them, separate from vocabulary breadth. In addition, we found that L2 English vocabulary breadth explained a substantial amount of variance in the measures.

These findings have both theoretical and pedagogical implications. To reiterate, we demonstrated that DM knowledge can be empirically separated from vocabulary breadth. In addition, the study shed light on the extent to which vocabulary breadth can be involved in DM measures. This is an obvious benefit for future research into the construct of L2 DM knowledge.

Above all, however, the findings yield insights into measuring and assessing DM knowledge. Recognising DM knowledge as a separate construct whose measures, nevertheless, also elicit other kinds of vocabulary knowledge, suggests that for practical purposes, controlling for that latter kind of knowledge can be useful in research measuring L2 DM knowledge. In general, the varying correlations (Table 2) between our DM measures indicate that they did not measure exactly the same kind of DM knowledge (see also the EFA results in Table 5), which suggests that a number of different DM tasks are needed to test DM knowledge comprehensively.

Regarding pedagogical implications, our finding about the separability of DM knowledge from vocabulary breadth suggests that the two can be mastered somewhat differently and, thus, there can be some value in teaching DM in addition to teaching to increase learners' vocabulary size. This conclusion is also supported by Kieffer and Lesaux (2012), who found that morphological awareness is a significant contributor to learners' reading ability, and by Nation (2001), who built an argument for the value of knowing affixes in understanding derived words. The question, of course, remains whether there is value in teaching, learning, and assessing affixes as separate morphemes with their own meanings and syntactic functions, or whether whole-word derivational forms should be learned. This can be further explored in future research.

Finally, we would like to acknowledge the limitations of the study. The first is not a limitation as such but has to do with the particular set of measures in this study. Had we used different measures, which we touched upon with reference to the word segmentation measure, the results might have been somewhat different. Indeed, when interpreting the results, it should be remembered that the vocabulary breadth and DM knowledge factors were based on the performances of a particular sample of L2 learners on particular measures. Had we used a different set of measures, for example, a recently published instrument (Sasao & Webb, 2017) that measures receptive knowledge of form, meaning, and use of derivational affixes, as well as Webb et al.'s (2017; Iwaizumi & Webb, 2021) updated version of the VLT, different factors might have emerged. These instruments could have been viable additions or substitutions to the measures in the present study. However, our data were collected before these instruments were made available to the academic community. We also acknowledge that even if our sample was larger than in Spencer et al.'s (2015) study and close to the sample size in González-Fernández and Schmitt's (2020) study, the total number of participants in the study was somewhat small, even if satisfactory. It could be that our confirmatory models had trouble fitting to the data due to a lack of power. Furthermore, similar to Koizumi and In'nami (2020), some of the measures we used were multiple-choice, a format prone to guessing. Thus, further studies would be required to improve the confidence in these findings, including, for example, neurolinguistic research (e.g., Leminen et al., 2019).

Despite the challenges arising from the use of SEM, perhaps in relation to our sample size, and despite the limitations of the study, this study contributed to conceptualising derivational morphology as an aspect of word knowledge that interacts with vocabulary breadth. We, thus, hope the study produced interesting results with regard to the conceptualisation of DM knowledge and will inspire further research on this topic.

Acknowledgements

The data were collected as a part of the TOPLING project (project no. 131956) supported by Academy of Finland.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The Finnish data for the study were collected as part of the Topling project funded by the Academy of Finland, project number 131956.

ORCID iDs

Dmitri Leontjev https://orcid.org/0000-0003-0177-3681

Asko Tolvanen https://orcid.org/0000-0001-6430-8897

Supplemental material

Supplemental material for this article is available online.

Notes

1. The statistical figures in the Supplementary Materials are based on the whole sample, which included a small group of university employees which we excluded from the analyses in this paper. We note that the same generalisations can be made should the performance of this group be included in the models, that is, the derivational morphology (DM) knowledge construct appears to be separate from vocabulary breadth.
2. Based on the overall validity evidence of the measures (see Supplementary Materials), we kept all the measures in the analysis. However, we also run exploratory factor analysis (EFA) with just the measures that had higher than .8 (Cronbach) alpha reliability (i.e., the three Vocabulary Levels Test [VLT] measures and affix elicitation, non-word affix elicitation, and prefix elicitation tasks. The EFA analysis suggested that there were two factors, one onto which the three VLT measures loaded and one with the three DM measures. Namely, the fit of the one-factor model was very low, $\chi^2(9)=27.201$, $p=.001$, whereas the two-factor model fit the data well, $\chi^2(4)=1.416$, $p=.842$. The fit of the two-factor model was also significantly different from the one-factor model, $\chi^2(5)=24.617$, $p<.001$. The loadings of the DM measures to the corresponding factors ranged from about .6 to about .85.

References

Akande, A. (2003). Acquisition of the inflectional morphemes by Nigerian learners of English. *Nordic Journal of African Studies*, *12*(3), 310–326. http://www.njas.helsinki.fi/pdf-files/vol12num3/akande2.pdf

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253–279. https://doi.org/10.1093/ijl/6.4.253

Carlisle, J. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing*, *12*(3), 169–190. https://doi.org/10.1023/A:1008131926604

Child, D. (2006). *The essentials of factor analysis* (3rd ed.). Continuum.

Chuenjundaeng, J. (2006). *An investigation of SUT students' receptive knowledge of English noun suffixes* [Master's thesis, Suranaree University of Technology]. Suranaree University of Technology Intellectual Repository. http://sutir.sut.ac.th:8080/jspui/handle/123456789/1585

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. https://rm.coe.int/1680459f97

Dawson, N., Rastle, K., & Ricketts, J. (2021). Finding the man amongst many: A developmental perspective on mechanisms of morphological decomposition. *Cognition*, *211*, Article 104605. https://doi.org/10.1016/J.COGNITION.2021.104605

Friedline, B. (2011). *Challenges in the second language acquisition of derivational morphology: From theory to practice* [Doctoral dissertation, University of Pittsburgh]. D-Scholarship. http://d-scholarship.pitt.edu/id/eprint/8351

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Goodwin, A. P., Huggins, A. C., Carlo, M., Malabonga, V., Kenyon, D., Louguit, M., & August, D. (2012). Development and validation of extract the base: An English Derivational Morphology Test for third through fifth grade monolingual students and Spanish-speaking English language learners. *Language Testing*, *29*(2), 265–289. https://doi.org/10.1177/0265532211419827

Goodwin, A. P., Petscher, Y., Carlisle, J., & Mitchell, A. (2017). Exploring the dimensionality of morphological knowledge for adolescent readers. *Journal of Research on Reading*, *40*(1), 91–117. https://doi.org/10.1111/1467-9817.12064

Goodwin, A. P., Petscher, Y., & Tock, J. (2020). Morphological supports: Investigating differences in how morphological knowledge supports reading comprehension for middle school students with limited reading vocabulary. *Language, Speech, and Hearing Services in Schools*, *51*(3), 589–602. https://doi.org/10.1044/2020_LSHSS-19-00031

Goodwin, A. P., Petscher, Y., & Tock, J. (2021). Multidimensional morphological assessment for middle school students. *Journal of Research on Reading*, *44*(1), 70–89. https://doi.org/10.1111/1467-9817.12335

Goodwin, A. P., Petscher, Y., Tock, J., McFadden, S., Reynolds, D., Lantos, T., & Jones, S. (2020). Monster, P.I.: Validation evidence for an assessment of adolescent language that assesses vocabulary knowledge, morphological knowledge, and syntactical awareness. *Assessment for Effective Intervention*, *47*(2), 89–100. https://doi.org/10.1177/1534508420966383

Hayashi, Y., & Murphy, V. (2011). An investigation of morphological awareness in Japanese learners of English. *Language Learning Journal*, *39*(1), 105–120. https://doi.org/10.1080/09571731003663614

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303–317. https://doi.org/10.1017/S0272263199002089

Hu, L.-T., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly*, *8*(3), 250–276. https://doi.org/10.1080/15434303.2011.582203

Iwaizumi, E., & Webb, S. (2021). To what extent does productive derivational knowledge of adult L1 speakers and L2 learners at two educational levels differ? *TESOL Journal*, *12*(4), Article e640. https://doi.org/10.1002/TESJ.640

Iwaizumi, E., & Webb, S. (2022). Measuring L1 and L2 productive derivational knowledge: How many derivatives can L1 and L2 learners with differing vocabulary levels produce? *TESOL Quarterly*, *56*(1), 100–129. https://doi.org/10.1002/tesq.3035

Kieffer, M., & Lesaux, N. (2012). Direct and indirect roles of morphological awareness in the English reading comprehension of native English, Spanish, Filipino, and Vietnamese speakers. *Language Learning*, *62*(4), 1170–1204. https://doi.org/10.1111/j.1467-9922.2012.00722.x

Kline, R. (2011). *Principles and practice of structural equation modeling* (3rd ed.). The Guilford Press.

Koizumi, R., & In'nami, Y. (2020). Structural Equation Modeling of vocabulary size and depth using conventional and Bayesian methods. *Frontiers in Psychology*, *11*, Article 618. https://doi.org/10.3389/fpsyg.2020.00618

Leminen, A., Smolka, E., Duñabeitia, J. A., & Pliatsikas, C. (2019). Morphological processing in the brain: The good (inflection), the bad (derivation) and the ugly (compounding). *Cortex*, *116*, 4–44. https://doi.org/10.1016/j.cortex.2018.08.016

Leontjev, D., Huhta, A., & Mäntylä, K. (2016). Word derivational knowledge and writing proficiency: How do they link? *System*, *59*, 73–89. https://doi.org/10.1016/j.system.2016.03.013

Lieber, R. (2017). Derivational morphology. In *Oxford research encyclopedia of linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.248

Linacre, J. M. (2015). *Winsteps® Rasch measurement computer program* [Computer software]. Winsteps.com

Mäntylä, K., & Huhta, A. (2013). Knowledge of word parts. In J. Milton & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (pp. 45–59). Palgrave Macmillan.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Milton, J., & Fitzpatrick, T. (Eds.). (2013). *The dimensions of word knowledge*. Palgrave Macmillan.

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing*, *36*(1), 101–123. https://doi.org/10.1177/0265532217725776

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System*, *28*(2), 291–304. https://doi.org/10.1016/S0346-251X(00)00013-0

Muthén, L., & Muthén, B. (1998–2015). *Mplus user's guide* (7th ed.).

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of Exploratory Factor Analysis. *Language Learning*, *65*(Suppl. 1), 9–36. https://doi.org/10.1111/LANG.12111

Raubenheimer, J. (2004). An item selection procedure to maximise scale reliability and validity. *SA Journal of Industrial Psychology*, *30*(4), 59–64. https://doi.org/10.4102/sajip.v30i4.168

Richards, J. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*(1), 77–89. https://doi.org/10.2307/3585941

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin, 28*(12), 1629–1646. https://doi.org/10.1177/014616702237645

Sasao, Y., & Webb, S. (2017). The word part levels test. *Language Teaching Research*, *21*(1), 12–30. https://doi.org/10.1177/1362168815586083

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, *48*(2), 281–317. https://doi.org/10.1111/1467-9922.00042

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition*, *19*(1), 17–36. https://doi.org/10.1017/S0272263197001022

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1177/026553220101800103

Schmitt, N., & Zimmerman, C. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*(2), 145–171. https://doi.org/10.2307/3588328

Seidenberg, M., & Gonnerman, L. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, *4*(9), 353–361. https://doi.org/10.1016/S1364-6613(00)01515-1

Spencer, M., Muse, A., Wagner, R., Foorman, B., Petscher, Y., Schatschneider, C., Tighe, E. L., & Bishop, M. (2015). Examining the underlying dimensions of morphological awareness and vocabulary knowledge. *Reading and Writing*, *28*(7), 959–988. https://doi.org/10.1007/s11145-015-9557-0

Taft, M. (1981). Prefix stripping revisited. *Journal of Verbal Learning and Verbal Behavior*, *20*(3), 289–297. https://doi.org/10.1016/S0022-5371(81)90439-4

Tarka, P. (2018). An overview of structural equation modeling: Its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity*, *52*(1), 313–354. https://doi.org/10.1007/s11135-017-0469-8

Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language*, *28*(6), 649–667. https://doi.org/10.1016/0749-596X(89)90002-8

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated vocabulary levels test. *ITL International Journal of Applied Linguistics*, *168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

Wolf, E., Harrington, K., Clark, S., & Miller, M. (2013). Sample size requirements for structural equation models. *Educational and Psychological Measurement*, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Appendix 1

Standardised coefficients in the two-factor CFA model.

	Estimate	95% CIs		*p*-value*
		Lower	Upper	
Vocabulary BY				
VLT 3000	.906	.864	.947	
VLT 5000	.887	.839	.935	
VLT AWL	.929	.893	.964	
DM BY				
Affix elicitation	.767	.656	.878	
Non-word aff. elicitation	.792	.717	.868	
Pref. elicitation	.799	.721	.877	
Grammar recognition	.820	.754	.886	
Recognition of meaning	.798	.727	.868	
Free prod.	.678	.589	.767	
Metaling. Prompts	.679	.577	.781	
Word segmentation	.317	.126	.509	.001
DM WITH VOC	.799	.693	.905	
Prod WITH Met. prompts	.400	.240	.561	

Note: CI: confidence interval; VLT: Vocabulary Levels Test; AWL: Academic Word List; DM: derivational morphology; CFA: confirmatory factor analysis.

*$p < .001$ unless otherwise stated.