Familiarity breeds dissent: Reliability analyses for British-English idioms on measures of familiarity, meaning, literality, and decomposability

Emily Nordmann1, Alexandra A. Cleland1, & Rebecca Bull2

1School of Psychology, William Guild Building, University of Aberdeen, Aberdeen, UK, AB24 2UB

2National Institute of Education, Nanyang Technological University, 1 Nanyang Walk, Singapore, 637616

Address correspondence to:

Emily Nordmann

School of Psychology

William Guild Building

University of Aberdeen

Aberdeen

AB24 2UB

Email: emily.nordmann@abdn.ac.uk

Telephone: +44 (0) 1224 273646

Fax: +44 (0) 1224 273426

Abstract

To date, there have been several attempts made to build a database of normative data for English idiomatic expressions (e.g., Libben & Titone, 2008; Titone & Connine, 1994), however, there has been some discussion in the literature as to the validity and reliability of the data obtained, particularly for decomposability ratings. Our work aimed to address these issues by looking at ratings from native and non-native speakers and to extend the deeper investigation and analysis of decomposability to other aspects of idiomatic expressions, namely familiarly, meaning and literality. Poor reliability was observed on all types of ratings, suggesting that rather than decomposability being a special case, individual variability plays a large role in how participants rate idiomatic phrases in general. Ratings from native and non-native speakers were positively correlated and an analysis of covariance found that once familiarity with an idiom was accounted for, most of the differences between native and non-native ratings were not significant. Overall, the results suggest that individual experience with idioms plays an important role in how they are perceived and this should be taken into account when selecting stimuli for experimental studies. Furthermore, the results are suggestive of the inability of speakers to inhibit the figurative meanings for idioms that they are highly familiar with.

*Keywords:* idioms, familiarity, reliability

1. Introduction

Idioms are expressions that are more than the sum of their parts. The figurative meaning of an idiom can be difficult to ascertain from a compositional analysis of the literal meaning of the component words (e.g., *kick the bucket* literally means *to strike a container with one’s foot* and figuratively means *to die*). Much of the research to date has focused upon how, and if, both of these meanings are activated during comprehension and production, and how the lexical representation of idioms can be incorporated into existing models of language. Item generation for psycholinguistic research can often be an arduous process given the large number of characteristics that even single words can be matched upon. Idiomatic expressions provide the additional concepts of decomposability and literality and consequently designing well-matched experimental stimuli can become very difficult. The purpose of the current work was not to add to the body of literature concerned with the lexical representation of idioms, but rather to investigate the methods used to obtain normative data that is then used to select experimental stimuli.

Whilst there are large databases of information about single words (e.g., CELEX (Baayen, Piepenbrock, & Van Rijn, 1993) and SUBTLEX (Brysbaert & New, 2009)), normative data for idioms exists in the form of a handful of published papers (e.g., Bonin, Méot & Bugaiska, 2013; Caillies, 2009; Libben & Titone, 2008; Tabossi, Ardunio, & Fanari, 2011; Titone & Connine, 1994) that cover several hundred expressions in different languages. Whilst the paucity of data is unfortunate, what is of greater concern is the lack of reliability reported for these data. In particular, concerns have been raised over the relevance of the construct of semantic decomposability given that studies show few consistent effects of decomposability (Tabossi, Wolf, & Koterle, 2009). Whilst the bulk of the discussion concerning the reliability of ratings has focused upon the aspect of decomposability, it is also worth considering the reliability of ratings obtained for other features of idiomatic expressions to determine whether or not decomposability represents a special case. The initial focus of the current work was therefore to investigate whether the concepts of familiarity, meaning and literality are subject to the same level of inter-rater variability.

*1.1 Familiarity*

Familiarity is considered to be a measure of the frequency with which a speaker or listener encounters a word or idiomatic expression (Gernsbacher, 1984). Titone and Connine (1994) describe familiarity as the “subjective frequency” (p255) of an idiom which can be viewed as distinct from ‘objective’ frequencies obtained from word counts in, for example, newspapers. Aside from the current lack of an objective frequency corpus for idioms, the inherent variability of individual experience with such expressions means that a subjective rating of familiarity is more likely to provide useful information. Familiarity also incorporates how well the meaning of a phrase is known or understood, and all norming studies to date have found strong positive correlations between familiarity and meaning judgements (e.g., Libben & Titone, 2008; Tabossi et al., 2011; Titone & Connine, 1994).

*1.2 Decomposability*

Decomposability is generally defined as whether the literal meanings of the component words of an idiomatic expression have a connection to the overall figurative meaning. A normally decomposable idiom is defined as one where the component words of a phrase are connected to the overall figurative meaning in a literal way, e.g., *play the market*. Abnormally-decomposable idioms are those items whereby the relationship between the component words and the figurative meaning is more metaphorical, e.g., *save your skin*. Non-decomposable idioms are those phrases that are seen as most stereotypically idiomatic, whereby the literal meaning of the component words bear no relation to the overall figurative meaning, e.g., *kick the bucket.* From a theoretical perspective, being able to manipulate the decomposability of stimuli for behavioural studies is important as it allows conclusions to be drawn about the lexical representation of idiomatic expressions. If decomposability could be demonstrated to affect production latencies then this would be strong evidence for a special type of idiom representation where such information could be encoded. Conversely, if decomposability appears to have no effect on language processing then this would support research that claims idioms utilise the same levels of lexical representation and processing mechanisms that literal sentences do (e.g., Konopka & Bock, 2009). So far the experimental evidence of an effect of decomposability has been mixed (e.g., Cutting & Bock, 1997; Konopka & Bock, 2009; cf. Sprenger, Levelt & Kempen , 2006.; Nordmann et al., 2013) and this lack of consensus may be a result of the data used to group stimuli as decomposable or non-decomposable.

*1.2 Literality*

The defining feature of an idiomatic expression is the figurative meaning the phrase carries, regardless of its relationship to the literal meaning of the component words. However, independently of decomposability it is possible to judge idioms on whether or not they have the potential for a literal interpretation. For example, *kick the bucket* has a well-formed literal meaning, that is, *to strike a container with one’s foot*. This literal interpretation is available despite the fact that as an idiomatic expression with the figurative meaning *to die*, *kick the bucket* is highly semantically non-decomposable. In contrast, it is difficult to imagine how one would literally *talk a mile a minute* even though the idiomatic meaning of *to speak quickly* is decomposable.

In terms of norming studies, Libben and Titone (2008) and Tabossi et al. (2011) both found that participants’ literality ratings correlated negatively with familiarity ratings (higher literality rating correlated with lower familiarity), although Titone and Connine (1994) found no such relationship. The link between literality and decomposability is also unclear. Titone and Connine found literality to be negatively correlated with abnormal decomposability, that is, those expressions whose component words were only metaphorically related to their figurative meaning were less likely to have a well-formed literal meaning. In contrast, Libben and Titone found that literality was negatively correlated with all measures of decomposability (global decomposability, normal, and abnormal decomposability) whilst Tabossi et al. (2011) found no relationship between literality and any of the decomposability measures.

*1.4 Reliability analyses*

A number of studies (Gibbs & Nayak, 1989; Libben & Titone, 2008; Tabossi, Fanari, & Wolf, 2008; Titone & Connine, 1994) have asked participants to make categorical judgements regarding decomposability and, if it was indicated that an item was thought to be decomposable, they were then asked to make a further categorical decision between normal and abnormal decomposability. Those studies that have used categorical methods of idiom classification have generally found low levels of agreement. Titone and Connine (1994) regarded an idiom as having been reliably categorised if agreement between participants reached 67% according to an approximation of the binomial distribution. With this criterion they found that only 40% of the idioms used were reliably classified as being non-decomposable, normally decomposable, or abnormally decomposable. These results stand in stark contrast to Gibbs and Nayak (1989) who showed agreement levels above 75% for 37 out of 40 idiomatic expressions. However, as noted in Tabossi et al. (2008), the idioms selected for use in Gibbs and Nayak were not chosen at random but selected by the authors as likely to be non-decomposable, normally decomposable, or abnormally decomposable. Therefore, it is unclear whether the high agreement levels observed by Gibbs and Nayak reflect participants’ true ability to reliably categorise idioms, or if the pre-selection of those items that were most likely to strongly fit into one of the categories influenced the results.

A primary concern of our work regards the use of categorical ratings for decomposability. Gibbs and Nayak (1989) suggest that decomposability exists on a continuum rather than there being a precise definition that can be used to sort idioms into dichotomous categories. The question raised is whether the low agreement ratings observed are a product of the categorical rating system, when a continuous scale, for example a Likert scale, would be more appropriate to capture the nuances of decomposability.

Tabossi et al. (2008) and Tabossi et al. (2011) utilised a 7-point Likert scale in order to obtain decomposability ratings. Those idioms that were rated as having a mean score of four or less were categorised as non-decomposable. The results from both studies indicated that the use of a Likert scale to rate decomposability does not appear to increase agreement between participants. Tabossi et al. (2008) reported that for a subset of 16 idioms agreement was high (83.54%). However, for the remaining 64 expressions agreement was on average 56.14%, a proportion not reliably different from chance. Similarly, Tabossi et al. (2011) found that only 28% of their idiomatic expressions could be reliably categorised using a 67% agreement level. As a further measure of agreement, Tabossi et al. (2008) also provided the results of an inter-rater reliability analysis using Kendall’s coefficient of concordance. The authors raised concerns that the dichotomisation of idioms into decomposable and non-decomposable may have led to a lower estimate of reliability, however, the results were in line with the previous analyses and found poor reliability of .13.

One of the arguments for the use of Likert scales to assess decomposability is that such scales may be able to capture the nuances of idiom decomposability better than asking participants to make a dichotomous categorisation. However, in the literature to date, Likert scores are still then categorised according to a cut-off point (e.g., in Tabossi et al. (2008) a mean rating of four or less is considered non-decomposable). By subsequently categorising the Likert scores the subtlety of the scale is lost, particularly for those items that are not considered at the extreme ends of the scale. For example, using a cut-off of a mean rating of four or less to determine decomposability category, if an idiom was rated by 50% of participants as decomposable (more than 4) and by 50% as non-decomposable (less than 4), this item would not be considered to have been reliably categorised. However, this analysis does not allow for the possibility that 100% of scores may lie, for example, between 3.5 and 4.5 on the Likert scale. By using percent agreement, those idioms that are not strongly decomposable or non-decomposable become classified as unreliable. Hayes and Krippendorff (2007) suggest that the statistic of choice for inter-rater reliability should be Krippendorff’s alpha. Nordmann et al. (2013) used Krippendorff’s alpha to judge the reliability of decomposability ratings obtained for 30 experimental items made using a three point Likert scale and found moderate reliability (*α*= .41) The current work, therefore, seeks to extend the use of this measure of reliability for idiom norming data. This measure can be used with nominal, ordinal, interval and ratio data, as well as with unequal numbers of participants and missing data. These features allow direct comparisons to be made between studies that utilise different rating scales and unequal numbers of participants.

Our first aim was to determine if any of the results from previous studies could be replicated. Our second aim was to apply the same reliability analyses using Krippendorff’s alpha to the measures of familiarity, meaning, literality, and decomposability in order to determine if participants are able to reliably classify idioms on any of the features identified by psycholinguistic research as important. We hypothesised that the reliability of participant ratings of familiarity and literality would be moderately strong given that familiarity should show something akin to a word frequency effect and literality judgements can be made objectively.

*Native vs. non-native speakers*

Research concerning idioms and non-native speakers has predominantly focused upon how non-native speakers learn idioms in their second language (L2), the problems they face, and how such expressions may be represented in the lexicon (e.g., Cooper, 1999; Liontas, 2002), however, Abel (2003) reports what appears to be the only norming study of idioms using non-native speakers. Whilst native speakers appear to show a bias towards rating idioms as non-decomposable, the results demonstrated a tendency for non-native speakers to categorise idioms as decomposable with 56.5% rated as such. The study used dichotomous categories for decomposability and, with an agreement criterion of 75%, found a total of 48.5% of idioms (study 1) were reliably sorted into either category; a level slightly higher than the 40% reported by Titone and Connine (1994) although still well below the 92.5% of Gibbs and Nayak (1989). The results also showed an effect of decomposability on familiarity, with decomposable idioms rated as more familiar, with a mean familiarity rating of 4.90 compared to only 2.99 for non-decomposable idioms. For native speakers (taken from the Titone & Connine dataset), although decomposable idioms were rated as more familiar than non-decomposable idioms the difference was much less pronounced (M = 5.92 for decomposable idioms, M = 5.76 for non-decomposable idioms) suggesting that the ability to semantically decompose an idiom makes it easier for non-native speakers to comprehend and use it in their L2.

The results of Abel (2003) provide a solid starting point, but there are further analyses that need to be conducted in order to make the comparison between native and non-native speakers’ ratings of increased use to idiom research. First, it would be useful to obtain correlations between the measures (familiarity, literality, decomposability, etc.) as reported in native norming studies, as well as correlations between native and non-native speakers. Secondly, by carrying out reliability analyses on non-native speakers’ ratings and comparing the level of agreement to that of native speakers, it should be possible to determine if the lack of reliability observed in native ratings is due to experience with the idiom.

2. Method

*2.1 Participants*

Seventy-six participants (40 females) were recruited from the University of Aberdeen’s School of Psychology as part of a course requirement. Participants had an average age of 20.71 (3.27 = SD). Forty-four participants were native speakers of English, 32 were fluent non-native speakers. All non-native speakers were enrolled on a degree programme and therefore had passed the IELTS (International English Language Testing System) proficiency exam. There was a variety of first languages spoken by non-native speakers, the majority being Eastern-European languages. Participants were awarded course credit for taking part in the survey.

*2.2 Materials*

Each participant was given a survey that asked them to rate 100 idioms (see appendix for full list of idioms) on four different criteria: how familiar the idiom was to them; how well they thought they knew the meaning; whether they thought the idiom had a plausible literal interpretation; and, if the idiom was decomposable. The instructions for each rating type were based on those in Titone and Connine (1994). There were two versions of the decomposability rating section. Forty-one participants (16 non-native) were asked to choose whether they thought the idiom was decomposable or non-decomposable. Following Titone and Connine, if they indicated the idiom was decomposable they were then asked if they thought it was normally or abnormally decomposable. The remaining 35 participants (16 non-natives) were asked to rate on a seven-point Likert scale how decomposable they thought the idiom was.

The length of each expression ranged from 2 – 7 words, with an average length of 3.71 words. Unlike Libben and Titone (2008) and Abel (2003), but following Titone and Connine (1994), there were no restrictions on the syntactic structure of the idioms. The items were presented in the same fixed random order for all participants.

*2.3 Procedure*

Participants were given one hour to complete the survey which was administered as a paper and pencil test. The survey contained an explanation of what each rating meant and provided examples for the literality and decomposability sections. The instructions were taken from those provided in Titone and Connine (1994). The ratings were also explained verbally before participants began the survey. All participants were given the rating sections in the same order; familiarity, meaning, literality, decomposability.

3. Results

See appendix for ratings from native speakers. The full set of rating data from both native and non-native speakers is freely available at <http://tinyurl.com/nativespeakers> and <http://tinyurl.com/nonnativespeakers>.

*3.1 Reliability*

Due to a printing error, categorical decomposability ratings were not obtained for one item, leaving a total of 99 idioms in the dataset. Analyses of the agreement levels for the categorical decomposability rating from native speakers were conducted using the same method as described in Titone and Connine (1994). An approximation of the binomial distribution for the decomposability ratings indicated that a 72.5% agreement rate (that is, 29/40 participants categorising an idiom as decomposable or non-decomposable) would be significant at *p* < .01 and an agreement rate of 67.5% would be significant at *p* < .05, similar to the values used by Titone and Connine. Using the 67.5% criterion in order to maximise the number of idioms that could be categorised, a total of 49 idioms (49.5%) were reliably categorised as non-decomposable with 12 (12.12%) categorised as decomposable. Of these 12 idioms, five (5.05% of total number of idioms) were reliably categorised as normally decomposable and one (1.01%) was categorised as abnormally decomposable. For the non-decomposable category Titone and Connine reported that 42% of idioms were reliably categorised, with 22% of idioms reliably categorised as decomposable, a percentage far higher than found in the current study.

For non-native speakers, the 67.5% agreement level resulted in a total of 48 (48.48%) idioms being reliably categorised as non-decomposable and 12 (12.12%) as decomposable. Of those 12, five (5.56%) were categorised as normally decomposable and two (2.02%) as abnormally decomposable. Unlike the results of Abel (2003), non-native speakers showed the same bias towards reliably categorising idioms as non-decomposable as native speakers, similar to the findings of Titone and Connine (1994).

The use of percent agreement to compare the results of different studies is problematic due to the non-random nature of the stimuli. In order to present a more robust measure of agreement that can be directly compared with other data, inter-rater reliability was also calculated. The first set of analyses aimed to investigate whether participants were able to reliably rate the idioms on the four measures by calculating Krippendorff’s alpha (Hayes & Krippendorff, 2007). Krippendorff’s alpha is computed by calculating disagreements instead of correcting percent-agreements and Hayes and Krippendorff provide a freely-available macro at <http://tinyurl.com/834n94y>. For all measures, inter-rater reliability was poor for both native and non-native speakers (see Table 1) although agreement was higher for native speakers.

Previous research has suggested that the lack of reliability in participants’ ratings of idiom decomposability is indicative of a core issue regarding the validity of decomposability as a psychologically valid construct that can affect the production and comprehension of idioms. Our results clearly support previous evidence suggesting participants are unable to give reliable ratings.

Table 1. *Krippendorff’s alpha inter-rater reliability for native and non-native speakers.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Familiarity** | **Meaning** | **Literality** | **DC**  **(Likert)** | **DC**  **(cat)** | **Norm DC** |
| **Native** | .27 | .24 | .21 | .26 | .16 | .10 |
| **Non-native** | -.02 | .14 | .12 | .12 | .20 | .08 |

*Note.* DC (Likert) = decomposable on Likert scale, DC (Cat) = decomposable on categorical ratings, Norm DC = normally decomposable.

Decomposability is a relatively abstract characteristic and it could be argued that each participant interprets the concept of decomposability in a different way, assigning differential semantic weight to the individual components of idiomatic phrases, thus leading to the lack of agreement in rating studies. If this is the case, then one is able to argue that decomposability as a variable in psycholinguistic research is problematic, as it is likely to be individual, rather than lexical representational differences that guide the results. However, reliability was also poor for familiarity, meaning, and literality. It is more difficult to make the case that these concepts are as abstract and as susceptible to individual variability as decomposability. Although one would expect individual variability to play a role in ratings of familiarity and meaning, particularly for non-native speakers, one would also expect something akin to a word frequency effect with some level of agreement around those items that are more or less common in every-day speech.

Whether or not a phrase has a literal interpretation should be minimally affected by individual variability and is most comparable to the notion of decomposability. The binominal analysis applied to decomposability was also conducted on the results of the literality scale, with a score of four or less indicating an item that was low in literality and a score of more than four indicating an idiom with a highly-literal interpretation. Using the 67.5% agreement criterion, only 20 idioms could be reliably classified as high in literality and 35 as low. Agreement ratings were therefore lower for literality than for decomposability in line with the results from the inter-rater reliability analysis. The results suggest that the lack of reliability in participant ratings is not unique to the concept of decomposability and may reflect a wider issue concerning individual variability in idiom comprehension and perception.

*3.2 Native vs. non-native speakers*

Analyses were then conducted to determine the differences in ratings between native and non-native speakers on the four measures. It was hypothesised that there would be significant differences in both familiarity and meaning, with native speakers rating all idioms more highly on both measures. For literality and decomposability it was hypothesised that there should be no significant difference in the ratings. Whilst the former measures are dependent upon prior experience it should be possible for non-native speakers to rate fixed expressions on literality and decomposability objectively as would native speakers, assuming their language proficiency is high enough.

A one-way MANOVA[[1]](#footnote-1) looking at mean rating scores of familiarity, literality, decomposability and literality revealed a significant main effect of speaker (native vs. non-native) (Wilks λ = .443, *F* (8, 181) = 28.48, *p* < .001, ηρ² = .557). Univariate analyses revealed a significant main effect of speaker for familiarity, meaning, literality, decomposability on the Likert scale, and abnormal decomposability (see Table 2). No effect of speaker was found for decomposability, non-decomposability, or normal decomposability on the categorical scale.

Table 2. *MANOVA results for comparisons between native and non-native speakers*.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **F** | **Sig.** | **Partial eta squared** | **Mean Native**  **(SD)** | **Mean non-native**  **(SD)** | |
| **Familiarity** | 99.497 | < .001 | .346 | 4.05  (1.21) | 2.49  (.90) |
| **Meaning** | 135.883 | < .001 | .420 | 4.87  (1.25) | 2.96  (.98) |
| **Literality** | 25.081 | < .001 | .118 | 4.14  (1.12) | 3.42  (.85) |
| **Decomposability**  **(Likert)** | 14.943 | < .001 | .074 | 3.62  (1.16) | 3.04  (.91) |
| **Decomposable (Categorical)** | .809 | .370 | .004 | .37  (.22) | .40  (.24) |
| **Non-decomposable** | .433 | .511 | .002 | .63  (.22) | .61  (.23) |
| **Normally decomposable** | 2.828 | .094 | .015 | .42  (.27) | .35  (.23) |
| **Abnormally decomposable** | 6.506 | .012 | .033 | .55  (.27) | .65  (.28) |

*Note*. Ratings for familiarity, meaning, literality, and decomposability (Likert) were given on a scale of 1 – 7 with higher ratings reflecting, e.g., higher familiarity. The categorical ratings represent mean proportion. Standard deviation is given in parentheses below mean ratings.

The results support the hypothesis that native speakers would rate idioms as overall more familiar and more meaningful than non-native speakers. The finding that native speakers also rated the idioms as significantly more literal and more decomposable than non-native speakers was unexpected. To determine if familiarity with an idiomatic expression was affecting speakers’ ratings, a one-way MANCOVA was performed with a composite measure of familiarity used as the covariate. This composite was obtained by summing the ratings given for familiarity and meaning. Controlling for familiarity, a main effect of speaker was found (Wilks λ = .839, *F* (6, 182) = 5.803, *p* < .001, ηρ² = .161). However, univariate analyses revealed that once the effect of familiarity was controlled for there were only significant differences between native and non-native speakers for categorical decomposability (*F* (1, 187) = 23.212, *p* < .001, ηρ² = .110) and non-decomposability (*F* (1, 187) = 24.559, *p* < .001, ηρ² = .116). The estimated means revealed that non-native speakers were more likely to rate an idiom as decomposable (native mean = .294, non-native mean = .478) and native speakers were more likely to rate an idiom as non-decomposable (native mean = .710, non-native = .526). By controlling for familiarity there were no longer significant differences between native and non-native ratings on literality or the decomposability Likert scale.

*3.3 Correlations between native and non-native speakers’ ratings*

Correlations were also performed to assess if there was a relationship between native and non-native speakers’ ratings. Whilst there were significant differences in the mean ratings between native and non-native speakers significant positive correlations were found for all measures (see Table 3).

Table 3. *Spearman’s Rho correlations between native and non-native speakers on familiarity, meaning, literality and the various measures of decomposability.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Fam** | **Mean** | **Lit** | **DC Likert** | **DC**  **(Cat)** | **NDC** | **Norm DC** | **Abnorm DC** |
| .55\*\* | .55\*\* | .74\*\* | .76\*\* | .77\*\* | .77\*\* | .29\* | .26\* |

*Note.* Fam = Familiarity, Mean = Meaning, Lit = Literality, DC (Likert) = decomposable on Likert scale, DC (Cat) = decomposable on categorical ratings, NDC = Non-decomposable, Norm DC = normally decomposable, Abnorm DC = abnormally decomposable.  
\* *p* < .05, \*\* *p* < .01.

The correlation analysis indicates that although non-native speakers rate idioms as significantly less familiar, meaningful, literal, and decomposable than native speakers, there is still a strong relationship between the two sets of ratings for each item. That is, the correlation suggests that the same idioms are considered to be higher in familiarity, meaningfulness, literality, and decomposability by both native and non-native speakers. Interestingly, the correlations between literality and decomposability were stronger than those for the experience based ratings of familiarity and meaning.

*3.4 Relationships between the ratings – native speakers*

Following the literature, a correlation was performed between the mean ratings for each item on all of the measures; familiarity, meaning, literality, decomposability (Likert), decomposable (categorical), non-decomposable (categorical), normal decomposability, and abnormal decomposability (see Table 4). For native speakers, in line with Titone and Connine (1994), Libben and Titone (2008), and Tabossi et al. (2011), familiarity was positively correlated with meaningfulness. Significant positive correlations were also found for familiarity with literality and both decomposability ratings with a negative correlation between familiarity and non-decomposability.

Table 4. *Spearman’s Rho correlation matrix for familiarity, meaning, literality, decomposability (Likert), decomposability (categorical), non-decomposability (categorical), normally decomposable, and abnormally decomposable for native speakers.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Familiarity** | **Meaning** | **Literality** | **DC (Likert)** | **DC (Cat)** | **NDC** | **Norm DC** |
| **Meaning** | .94\*\* |  |  |  |  |  |  |
| **Literality** | .37\*\* | .40\*\* |  |  |  |  |  |
| **DC (Likert)** | .42\*\* | .48\*\* | .62\*\* |  |  |  |  |
| **DC (Cat)** | .35\*\* | .44\*\* | .45\*\* | .82\*\* |  |  |  |
| **NDC** | -.38\*\* | -.48\*\* | -.49\*\* | -.86\*\* | -1.0\*\* |  |  |
| **Norm DC** | .16 | .19 | .23\* | .45\*\* | .42\*\* | -.42\*\* |  |
| **Abnorm DC** | -.18 | -.15 | -.19 | -.30\*\* | -.25\* | .25\* | -.85\*\* |

*Note.* DC (Likert) = decomposable on Likert scale, DC (Cat) = decomposable on categorical ratings, NDC = Non-decomposable, Norm DC = normally decomposable, Abnorm DC = abnormally decomposable.  
\* *p* < .05, \*\* *p* < .01

As in Titone and Connine (1994), literality was found to be negatively correlated with non-decomposability. Significant positive correlations were found between literality and decomposability (on both scales) and normal decomposability ratings. The results presented in Titone and Connine demonstrate a weak association between literality and decomposability.

The significant correlations between the categorical measures of decomposability and normal/abnormal decomposability reflect the fact that only those items which were indicated to be decomposable were further rated as normal/abnormal and are therefore not of interest theoretically. The finding that decomposability on the Likert scale was positively correlated with normal decomposability and negatively correlated with abnormal decomposability suggests that the use of a Likert scale does not lose this distinction.

*3.5 Relationships between the ratings – non- native speakers*

For non-native speakers, the pattern of results was similar to that of native speakers (see Table 5). Familiarity was positively correlated with meaning, literality, and normal decomposability, and was negatively correlated with abnormal decomposability. These results taken together suggest that those idioms that are more literal and can be semantically decomposed are easier for non-native speakers to learn as they can use semantic information from the individual component words. Mirroring the findings from native speakers, literality was positively correlated with normal decomposability, and negatively correlated with abnormal decomposability. Meaning ratings were also positively correlated with normal decomposability, suggesting that non-native speakers are better able to learn the meaning of those idioms whose component words have a literal, rather than metaphorical, relationship with their figurative meaning.

Table 5. *Correlation matrix for familiarity, meaning, literality, decomposability (Likert), decomposability (categorical), non-decomposability (categorical), normally decomposable, and abnormally decomposable for non-native speakers.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Familiarity** | **Meaning** | **Literality** | **DC (Likert)** | **DC (Cat)** | **NDC** | **Norm DC** |
| **Meaning** | .90\*\* |  |  |  |  |  |  |
| **Literality** | .62\*\* | .67\*\* |  |  |  |  |  |
| **DC (Likert)** | .19 | .18 | .24\* |  |  |  |  |
| **DC (Cat)** | .07 | .04 | .15 | .68\*\* |  |  |  |
| **NDC** | -.06 | -.05 | -.20 | -.72\*\* | -.96\*\* |  |  |
| **Norm DC** | .23\* | .22\* | .23\* | .36\*\* | .31\*\* | -.29\*\* |  |
| **Abnorm DC** | -.24\* | -.22\* | -.23\* | -.36\* | -.34\*\* | .31\* | -.98\*\* |

*Note.* DC (Likert) = decomposable on Likert scale, DC (Cat) = decomposable on categorical ratings, NDC = Non-decomposable, Norm DC = normally decomposable, Abnorm DC = abnormally decomposable.  
\* *p* < .05, \*\* *p* < .01

4. Discussion

The use of Krippendorff’s alpha to compare inter-rater reliability allows the first direct reliability comparison of the four measures of familiarity, meaning, literality and decomposability to be performed. The results suggested that decomposability is not special when it comes to poor agreement amongst raters. Reliability was poor for all measures, for both native and non-native speakers.

For familiarity and meaning, individual variability may account for a large proportion of the variance in the ratings given to different expressions. That is, variability between participants appears to be more important than variability between items. As such, it may be difficult to rely on normative data in the same way that one can for word frequency using e.g., CELEX (Baayen et al., 1993). The results from non-native speakers support this argument as the reliability analyses demonstrated the lowest agreement for non-native speakers on familiarity. This group is likely to have extremely divergent experiences of idiomatic expressions, as their exposure to English and the use of English idioms will be dependent upon a number of factors, e.g., their native language, the type and number of English books and television shows they have watched, etc. This is not to argue that the information collected on familiarity and meaning in norming studies is of no practical use. However, given the lack of reliability observed, we suggest that participant ratings should be collected on the items used in each experiment so that familiarity information from normative databases can be verified against sample specific ratings. Due to the importance of familiarity and meaning upon idiom processing and production (e.g., Libben & Titone, 2008), it is essential that items are accurately classified as being of high or low familiarity.

The poor reliability ratings observed for literality were unexpected. It was predicted that whether a phrase has a possible literal interpretation would produce reasonable inter-rater reliability given that, as a concept, literality is less abstract than decomposability. The original premise of the study was built upon the assumption that the poor reliability scores for decomposability were an aberration caused either by something inherent in the construct of decomposability or in the way that decomposability was measured. One possible explanation for the lack of reliability in literality ratings can be found in the work of Ariel (2002), who suggested that literal and non-literal meanings cannot always be distinguished from each other. We argue that native speakers are unable to inhibit their knowledge of the meaning of the phrase when they are making literality and decomposability judgements. If a participant is highly familiar with the idiomatic meaning then their perception of the phrase will be that it is semantically acceptable and well-formed. When asked to make literality and decomposability judgements, speakers are unable to supress this knowledge and are biased towards perceiving the idiom to have a literal and compositional interpretation. Familiarity, in other words, makes there seem like there is meaning where there is none. Keysar and Bly (1995) make a similar argument; that once the meaning of an idiom has been learnt and makes sense, its relative transparency increases and they propose quite strongly that our intuitions about idioms are not independent of our knowledge of the idiom’s meaning. In this respect, it may be that whilst decomposability is a psychologically relevant concept, it is not particularly useful because when we learn the meaning of an idiom we lose the ability to judge compositionality. This theory is supported by the results of the correlation analysis that showed a positive correlation between familiarity and literality, and meaning and literality. The more familiar a speaker is with an idiomatic expression, and the more they are sure of its meaning, the harder it is to ignore our knowledge when making ratings that require this knowledge to be suppressed. The results of the MANCOVA also support this argument. When familiarity was not taken into account, it was found that non-native speakers rated idioms as significantly less literal, however, once familiarity was controlled for this difference disappeared.

No relationship was observed between familiarity and literality in Titone and Connine (1994), however, in both Libben and Titone (2008), and Tabossi et al., (2011) familiarity was negatively correlated with literality, the opposite pattern to the one reported in this paper. Tabossi et al. suggest that their findings indicate that, for more familiar idioms, the figurative meaning is stronger than the available literal meaning. Given that their study was conducted using Italian idioms, it is not possible to conduct a direct comparison. It may be that the differences in results are due to the different idioms used in each study. For our findings, it may be that those idioms that are more literal and more semantically decomposable are easier to understand and to memorise as they may be able to benefit from non-idiomatic processing, and therefore are used more frequently and become more familiar to participants. Our study also revealed a strong positive relationship between literality and decomposability and strengthens the argument that those idioms that can have a possible literal interpretation are seen as more semantically decomposable.

As a counter-argument to this suggestion, the correlation analysis in Libben and Titone found that literality ratings correlated negatively with all measures of decomposability. They argue that these findings are indicative of the idiosyncratic nature of idiomatic expressions in that decomposable idioms do not have to be literally plausible and literally plausible idioms do not have to be decomposable. Similar to the differences found between our study and Tabossi et al. (2011), it is difficult to ascertain whether the observed contradictions in the results can be attributed to the different lists of idioms studied or if the results reflect a deeper theoretical or methodological issue. Replicating the findings of Titone and Connine (1994), native speakers were more likely to reliably rate idioms as non-decomposable, with 49.5% rated as such. It has been argued by Abel (2003) that this reflects a bias in native speakers to regard idioms as non-decomposable wholes whereas non-native speakers are biased towards a compositional analysis. The results reported here found both native and non-native speakers demonstrated a tendency to categorise idioms as non-decomposable, therefore this explanation is slightly problematic. It is possible that the bias reflects a difference in task difficulty. Those idioms that are highly non-decomposable, for example *kick the bucket,* may be easily classified as semantically non-decomposable. However, for those idioms that are, for example, abnormally decomposable, e.g., *meet your maker,* the decision is less clear-cut and may lead to a lack of agreement between raters. It is also possible, as noted in previous sections of this work, that comparing the percentage of items that are categorised as decomposable across studies is not appropriate given the restricted non-random nature of the item sample. We would therefore advise against drawing theoretical conclusions based upon such analyses.

Clearly the results of the reliability analyses of both types of scales suggest that caution should be taken when attempting to group idioms as decomposable and non-decomposable. However, there is reason to argue that the use of a Likert scale should be adopted as the scale of choice for rating decomposability. First, in all norming studies Likert scales are used to obtain information about the other characteristics of idiomatic expressions; familiarity, meaning, and literality. By using a Likert scale to rate decomposability it is possible to directly compare agreement levels. Secondly, as suggested in the introduction, decomposability exists on a scale, rather than as a well-defined dichotomy. In order to fully understand the effects of decomposability upon idiom comprehension and production, researchers should treat decomposability as a continuous variable and include the decomposability ratings obtained as a covariate rather than comparing two distinct groups. Finally, the results of the MANCOVA, comparing ratings from native and non-native speakers, revealed that once familiarity and meaning were controlled for significant differences were only found for those ratings obtained through the categorical decomposability measure. Given that non-native speakers appear to be rating idioms in much the same way as native speakers, as evidenced by the significant positive correlations between the measures, we argue that the significant differences observed for the categorical scale can be attributed to the type of rating rather than any intrinsic difference in the perception of decomposability between speakers.

Sample size is often cited as a potential cause of poor reliability; however, it appears unlikely that this is the cause of the results found in the current work given the strong correlations and the replication of previous findings. Although the total number of participants involved in the study was smaller than those of previous research, unlike previous studies all participants in our study rated all items. Tabossi et al. (2011) used a total of 740 participants, but because each participant saw only a subset of the full stimuli list this equated to a minimum of 40 ratings per idiom/variable, an N comparable to the familiarity, meaning and literality ratings we obtained. It is worth reiterating that due to there being two decomposability scales split between native and non-native speakers, the sample size for decomposability for each group is smaller than the other dimensions and for previous work. As such the conclusions drawn from the decomposability analyses may be treated with slightly more caution. A strength of our work is that all idioms were rated on all components by all participants and by doing so we were able to replicate the findings of previous work that had a much larger total sample size. Given the effect individual experience appears to exert over the data collected in norming studies, correlations performed between ratings that use different participants are unlikely to reflect the true relationship between variables, and for this reason we propose that a within-subjects design should be used when possible.

Whilst we would argue that the design of our study is a strength, it is important to recognise potential weaknesses in the methodology. The order in which participants rated the various measures was fixed and this may have had some effect on the ratings given. Rating how well one knows the meaning of an idiom must precede rating decomposability (as the definition is provided), but the order of literality and decomposability ratings should have been counterbalanced to avoid thinking about the literal interpretation of an idiom influencing the perception of decomposability.

The final aim of the study was to compare the ratings given to idiomatic expressions from native and non-native speakers. The purpose of our study was not to examine the cognitive representation of idioms for non-native speakers – the methodology of a survey is unsuited to this purpose – but rather to determine what non-native speakers’ ratings could contribute to our knowledge of the normative data on idioms.

That non-native speakers’ ratings on all measures were positively correlated with native speakers suggests that non-natives were not judging the idioms randomly without some degree of insight into the properties that they were rating. The results of the MANCOVA strongly suggest that familiarity with an idiom and its figurative meaning is an important factor in how idioms are perceived. Nippold and Rudzinski (1993) and Nippold and Taylor (1995) revealed that children found high-familiarity idioms easier to produce and comprehend, and this is linked to the idea that it is experience with idiomatic expressions that affects how participants will judge them on measures of literality and decomposability. That familiarity plays such a large role may also help explain why the reliability analyses revealed poor agreement between raters. Idiom familiarity is likely to be extremely susceptible to individual variability - as reflected in the poorest reliability ratings coming from non-natives on measures of familiarity. Given that the other measures appear to be influenced by familiarity, there is likely to be a residual effect of individual variability. The consequence of this variability is that those judgements that should be able to be made objectively, i.e., literality and decomposability, are less reliable than would be predicted. Related to this, we found stronger correlations between native and non-native speakers for literality and decomposability than for familiarity and meaning. This pattern of results suggests that to some degree the ratings of decomposability and literality can be made objectively. That this pattern of agreement does not translate into high reliability is supportive of the theory that the variability in familiarity has a large area of effect. As a point of interest, we did not control for idioms that are present in more than one language, e.g., *add fuel to the fire* exists in Serbian as *add oil to the fire/dolivati ulje na vatru*. Future research may wish to determine what influence, if any, this has upon how participants perceive idioms and whether there would be any differences in the literality and decomposability ratings given to the same idiom in different languages.

The differences between the ratings from native and non-native speakers suggest experience plays a large role in how idioms are perceived. We did not have detailed enough data to look at whether any relationships existed between, for example, L2 ability or number of years learning English and the various ratings. Whilst one would expect there to be a strong positive correlation between L2 ability and familiarity, the predictions for the relationship between ability and decomposability and literality are less clear and would be an interesting line of research for future work to pursue.

Finally, the current study did not investigate predictability or age-of-acquisition (AoA). These omissions were due to practical constraints concerning the length of the task for participants. Titone and Connine (1994) found a positive correlation between familiarity and predictability and Bonin et al. (2013) a positive correlation between decomposability and predictability. On the basis of these findings we would hypothesise the same pattern of results for predictability in the current study as we have found for familiarity. There is no reason to believe that inter-rater reliability of predictability would be higher than any of the other dimensions, however, this should be tested in future work. Bonin et al. also report strong negative correlations between AoA and familiarity, again leading us to the same prediction for the current study for AoA as found for familiarity.

We would like to make it clear that we are not arguing that participant ratings of idioms have *no* consistency or that normative data for idioms is of *no* use. We have replicated many of the correlations between the features of idioms found in previous research and these correlations are far from marginal. In addition, we found strong correlations between native and non-native ratings suggesting some degree of test-retest reliability. Due to these significant findings our argument is perhaps not as clear as we would prefer, and we ask the reader to indulge us in an analogy with which they may be all too familiar. Student X hands in two essays, one on the comprehension of idioms and one on production. Lecturer A gives the production essay a first and the comprehension essay a high 2:1. Lecturer B gives the production essay a 2:2 and the comprehension essay a high 3rd class mark. The relationship between the marks for the different essay types is consistent between the markers, both rated the production essay as higher quality, but there is no guarantee of reliability for the same essay. What we therefore argue is this: the nature of the relationships between various features of idiomatic expressions appears to be relatively strong and, with minor exceptions, are replicated across the literature. We believe that the evidence for these relationships is strong enough to base theoretical models upon and that studies such as our own, and previous (and future) norming studies, are a useful addition to the literature and are helping to progress our understanding of the processing of idioms. However, the crucial finding of this work is that inter-rater reliability *for individual features* is extremely poor. This does not mean that normative data for idioms is of no use; on the contrary, we have used this data to select idioms for our own experimental work. What we are arguing is that in experimental work that aims to manipulate, e.g., decomposability, ratings should also be obtained from the participants that take part in the experiment. Behavioural studies have consistently failed to find effects of decomposability. The prevailing opinion is that the null-effects are due to the fact that decomposability is not encoded at any level of representation. This may be the case, but we are not yet in a position to rule out that the null-effects are not the product of a clash between what the researcher has classified as ‘highly decomposable’ and what the participant regards as decomposable.

There are several conclusions that can be drawn from the current study. First, it appears that reliability for decomposability ratings can only be marginally improved by the use of a Likert scale rather than a categorical rating system. That said, in order to capture the nuances of idiom decomposability, and to allow for greater comparisons to be made between measures, the use of a Likert scale should be adopted. It can also be concluded that poor reliability is not a special feature of decomposability and that this also applies to ratings of familiarity, meaning, and literality. The results from the comparisons of native and non-native speakers suggest that a speaker’s familiarity and knowledge of an idiom greatly affects how different characteristics of an idiom are perceived. To this end, whilst norming databases can be useful in pre-selecting idioms for use in experimental research, ratings should be gained from each specific sample in order that the effects of familiarity can be reduced.

References

Abel, B. (2003). English idioms in the first language and second language lexicon: A dual representation approach. *Second Language Research, 19,* 329-358.

Ariel, M. (2002). The demise of the unique concept of literal meaning. *Journal of Pragmatics, 34*, 361-402.

Bonin, P., Méot, A., & Bugaiska, A. (2013). Norms and comprehension times for 305 French idiomatic expressions. *Behaviour Research Methods, 45* (4), 1259-1271.

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). The CELEX lexical database. *CD- ROM* . Philidelphia, PA.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977-990.

Caillies, S. (2009). Descriptions de 300 expressions idiomatiques familiarité, connaissance de leur signification, plausibilité littérale, "décomposabilité" et "prédictibilité". *Année Psychologique, 109 (3)*, 463- 508.

Cooper, T.C. (1999). Processing of idioms by L2 learners of English. *TESOL Quarterly*, *33 (2)*, 233-262.

Gernsbacher, A. M. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General, 113*, 256-281.

Gibbs, R.W., & Nayak, N.P. (1989). Psycholinguistic studies of the syntactic behaviour of idioms. *Cognitive Psychology, 21*, 100-138.

Hayes, A.F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1 (1),* 77-89.

Konopka, A.E., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology, 58*, 68-101.

Libben, M. R., & Titone, D.A. (2008). The multidetermined nature of idiom processing. *Memory & Cognition, 36 (6)*, 1103-1121.

Liontas, J. (2002). Context and idiom understanding in second languages. *EUROSLA Yearbook, 2,* 155-185.

Nordmann, E., Cleland, A. A., & Bull, R. (2013). Cat got your tongue? Using the tip-of-the-tongue state to investigate fixed expressions. *Cognitive Science*, *37* (8), 1553-1564.

Nippold, M. A., & Rudzinski, M. (1993). Familiarity and transparency in idiom explanation: A developmental study of children and adolescents. *Journal of Speech and Hearing Research, 36*, 728-737.

Nippold, M.A., & Taylor, C.L. (1995). Idiom understanding in youth: Further examination of familiarity and transparency. *Journal of Speech and Hearing Research, 45,* 384-391.

Sprenger, S.A., Levelt, W. J .M., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language, 54*, 161-184.

Tabossi, P., Arduino, L., & Fanari, R. (2011). Descriptive norms for 245 Italian idiomatic expressions. *Behavior Research Methods, 43*, 110-123.

Tabossi, P., Fanari, R., & Wolf, K. (2008). Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 313-327.

Tabossi, P., Wolf, K., & Koterle, K.W. (2009). Idiom syntax: Idiosyncratic or principled? *Journal of Memory and Language, 61,* 77-96.

Titone, D., & Connine, C. (1994). Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbol, 9 (4)*, 247-270.

Appendix 1. *Mean ratings from native speakers for familiarity, meaning, literality, decomposability (likert and categorical) and normal decomposability. Likert scale ratings were made between 1-7 with a higher number reflecting, e.g., higher familiarity. Categorical ratings reflect the proportion of participants who rated the idiom as such. Categorical proportions are complementary, for example, a decomposable rating of 0.28 would equal a non-decomposable rating of 0.72.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Idiom** | **Familiarity** | **Meaning** | **Literality** | **Decomposable**  **Likert** | **Decomposable Categorical** | **Normally Decomposable** |
| A game of two halves | 3.79 | 4.14 | 4.65 | 3.24 | 0.28 | 0.57 |
| **Across the pond** | **4.86** | **5.60** | **5.50** | **4.75** | **0.76** | **0.61** |
| Add fuel to the fire | 4.79 | 6.19 | 6.38 | 5.31 | 0.68 | 0.50 |
| **All talk and no trousers** | **3.28** | **5.55** | **4.55** | **3.94** | **0.48** | **0.33** |
| An Englishman's home is his castle | 2.36 | 3.74 | 4.83 | 4.12 | 0.40 | 0.40 |
| **Argue the Toss** | **1.76** | **3.21** | **2.77** | **3.13** | **0.44** | **0.60** |
| Armed to the teeth | 2.79 | 4.26 | 3.77 | 4.65 | 0.68 | 0.44 |
| **Asleep at the wheel** | **4.28** | **4.86** | **5.54** | **5.13** | **0.52** | **0.46** |
| At a loose end | 5.05 | 5.65 | 3.95 | 3.41 | 0.20 | 0.00 |
| **At the end of your tether** | **4.51** | **5.33** | **4.05** | **4.65** | **0.40** | **0.67** |
| Away with the fairies | 5.51 | 6.26 | 4.70 | 4.65 | 0.52 | 0.23 |
| **Beat around the bush** | **5.28** | **5.88** | **4.40** | **2.76** | **0.20** | **0.33** |
| Beat to the punch | 3.40 | 4.49 | 4.60 | 5.00 | 0.68 | 0.44 |
| **Been in the wars** | **4.88** | **5.81** | **6.40** | **5.56** | **0.76** | **0.67** |
| Bent out of shape | 4.16 | 4.90 | 5.05 | 3.53 | 0.24 | 0.17 |
| **Idiom** | **Familiarity** | **Meaning** | **Literality** | **Decomposable**  **Likert** | **Decomposable Categorical** | **Normally Decomposable** |
| **Birds of a feather** | **3.30** | **4.17** | **3.23** | **2.94** | **0.20** | **0.00** |
| Blow away the cobwebs | 3.66 | 4.75 | 5.41 | 4.35 | 0.48 | 0.25 |
| **Bob's your uncle** | **4.88** | **5.47** | **4.00** | **1.63** | **0.00** | **0.00** |
| Box clever | 2.28 | 3.00 | 2.56 | 3.47 | 0.36 | 0.11 |
| **Brass neck** | **2.69** | **3.23** | **3.21** | **1.69** | **0.16** | **0.25** |
| By a long chalk | 2.20 | 3.42 | 2.95 | 2.53 | 0.08 | 0.00 |
| **Cheap as chips** | **4.86** | **6.16** | **5.02** | **4.63** | **0.64** | **0.81** |
| Chinese whispers | 5.42 | 6.12 | 4.29 | 3.75 | 0.44 | 0.40 |
| **Chomping on the bit** | **3.12** | **4.00** | **4.42** | **2.72** | **0.12** | **0.00** |
| Coals to Newcastle | 1.37 | 1.67 | 2.98 | 1.88 | 0.04 | 1.00 |
| **Come a cropper** | **2.63** | **2.98** | **2.40** | **2.24** | **0.08** | **1.00** |
| Come up smelling of roses | 4.57 | 5.30 | 4.74 | 3.12 | 0.20 | 0.20 |
| **Daft as a brush** | **4.12** | **5.77** | **2.35** | **2.88** | **0.48** | **0.17** |
| Damp squib | 2.67 | 2.79 | 2.67 | 2.00 | 0.08 | 0.00 |
| **Death warmed up** | **4.47** | **5.19** | **3.05** | **4.53** | **0.44** | **0.30** |
| Do a runner | 5.53 | 6.14 | 4.60 | 5.63 | 0.64 | 0.80 |
| **Double Dutch** | **3.60** | **4.26** | **2.40** | **2.35** | **0.12** | **0.33** |
| Down in the dumps | 5.60 | 6.33 | 4.37 | 4.18 | 0.60 | 0.54 |
| **Idiom** | **Familiarity** | **Meaning** | **Literality** | **Decomposable**  **Likert** | **Decomposable Categorical** | **Normally Decomposable** |
| **Dunkirk spirit** | **1.58** | **2.53** | **3.33** | **2.53** | **0.24** | **0.17** |
| Fair crack of the whip | 4.07 | 4.81 | 5.21 | 3.65 | 0.40 | 0.60 |
| **Fall off the back of a lorry** | **3.22** | **3.83** | **5.56** | **4.18** | **0.32** | **0.63** |
| Fish out of water | 5.21 | 6.37 | 6.45 | 5.00 | 0.60 | 0.50 |
| **Fit as a fiddle** | **5.47** | **6.35** | **2.60** | **3.13** | **0.56** | **0.21** |
| Flash in the pan | 2.72 | 3.14 | 3.44 | 3.12 | 0.28 | 0.14 |
| **Flogging a dead horse** | **3.56** | **4.05** | **4.37** | **4.24** | **0.28** | **0.17** |
| Fly on the wall | 5.00 | 5.77 | 5.74 | 4.29 | 0.44 | 0.27 |
| **Foam at the mouth** | **4.90** | **5.26** | **5.67** | **4.19** | **0.44** | **0.36** |
| For donkey's years | 5.24 | 5.98 | 3.40 | 3.94 | 0.52 | 0.31 |
| **Full monty** | **4.42** | **5.64** | **2.86** | **4.00** | **0.40** | **0.67** |
| Get out of hand | 6.26 | 6.70 | 4.30 | 4.94 | 0.52 | 0.62 |
| **Get the nod** | **3.23** | **3.77** | **3.98** | **5.00** | **0.68** | **0.94** |
| Give someone stick | 4.65 | 5.84 | 3.93 | 3.22 | 0.16 | 0.25 |
| **Go down like a lead balloon** | **3.81** | **5.44** | **4.71** | **4.71** | **0.40** | **0.38** |
| Go off on one | 5.50 | 5.98 | 3.20 | 3.63 | 0.24 | 0.67 |
| **Idiom** | **Familiarity** | **Meaning** | **Literality** | **Decomposable**  **Likert** | **Decomposable Categorical** | **Normally Decomposable** |
| **Go spare** | **3.37** | **4.65** | **2.23** | **2.18** | **0.08** | **0.00** |
| Gone pear shaped | 5.02 | 6.07 | 3.67 | 2.38 | 0.20 | 0.20 |
| **Green fingers** | **4.56** | **5.26** | **4.21** | **4.06** | **0.56** | **0.43** |
| Hair of the dog | 4.63 | 4.49 | 3.98 | 1.24 | 0.00 | 0.00 |
| **In rude health** | **1.60** | **2.91** | **2.44** | **2.69** | **0.48** | **0.50** |
| In spades | 2.07 | 2.79 | 2.14 | 2.44 | 0.12 | 0.67 |
| **In the clink** | **2.21** | **2.77** | **2.60** | **3.33** | **0.32** | **0.50** |
| Keep your chin up | 6.09 | 6.76 | 6.29 | 5.41 | 0.60 | 0.71 |
| **Keep your wig on** | **4.72** | **6.14** | **6.14** | **3.11** | **0.32** | **0.29** |
| Lay down the law | 4.88 | 5.93 | 4.49 | 5.76 | 0.64 | 0.88 |
| **Like a bear with a sore head** | **2.28** | **3.51** | **4.81** | **3.29** | **0.32** | **0.29** |
| Lose your bottle | 3.63 | 5.21 | 4.81 | 3.29 | 0.52 | 0.25 |
| **Luck of the draw** | **4.72** | **6.02** | **4.63** | **5.24** | **0.68** | **0.81** |
| Nip in the bud | 4.41 | 5.60 | 4.12 | 3.35 | 0.40 | 0.44 |
| **Not give a monkey's** | **5.09** | **6.23** | **3.07** | **2.29** | **0.36** | **0.25** |
| Off your rocker | 5.44 | 6.33 | 3.65 | 3.18 | 0.28 | 1.00 |
| **On the blink** | **3.77** | **4.49** | **2.86** | **2.88** | **0.20** | **0.50** |
| **Idiom** | **Familiarity** | **Meaning** | **Literality** | **Decomposable**  **Likert** | **Decomposable Categorical** | **Normally Decomposable** |
| On the blower | 2.84 | 3.90 | 2.44 | 2.18 | 0.20 | 0.25 |
| **On the game** | **4.88** | **5.00** | **3.65** | **2.76** | **0.28** | **0.43** |
| On the take | 2.42 | 2.93 | 2.95 | 3.76 | 0.24 | 0.33 |
| **On the trot** | **3.81** | **3.95** | **3.24** | **2.88** | **0.04** | **0.00** |
| Out in the sticks | 3.70 | 4.35 | 3.07 | 3.94 | 0.24 | 0.33 |
| **Out of the blue** | **6.07** | **6.60** | **2.81** | **3.24** | **0.16** | **0.33** |
| Paper over the cracks | 2.67 | 3.88 | 4.51 | 2.80 | 0.36 | 0.56 |
| **Pardon my French** | **5.74** | **6.23** | **3.67** | **2.41** | **0.32** | **0.29** |
| Pull your finger out | 5.05 | 5.67 | 5.14 | 3.56 | 0.36 | 0.44 |
| **Quick off the mark** | **4.88** | **5.84** | **5.16** | **6.41** | **0.84** | **0.80** |
| Rake over old coals | 2.21 | 3.19 | 4.70 | 3.69 | 0.40 | 0.30 |
| **Rise to the bait** | **4.34** | **5.36** | **4.26** | **4.24** | **0.48** | **0.55** |
| Run out of steam | 5.09 | 6.29 | 5.58 | 5.83 | 0.84 | 0.45 |
| **Rushed off your feet** | **5.40** | **6.21** | **4.38** | **5.38** | **0.76** | **0.72** |
| Said the actress to the bishop | 2.35 | 1.95 | 3.59 | 1.72 | 0.08 | 0.50 |
| **Salt of the earth** | **3.19** | **3.65** | **3.74** | **2.59** | **0.08** | **0.50** |
| Saved by the bell | 5.59 | 6.49 | 4.67 | 5.53 | 0.76 | 0.58 |
| **Idiom** | **Familiarity** | **Meaning** | **Literality** | **Decomposable**  **Likert** | **Decomposable Categorical** | **Normally Decomposable** |
| **Send someone to Coventry** | **1.74** | **2.09** | **4.48** | **1.65** | **0.12** | **0.67** |
| Sign of the times | 3.26 | 4.39 | 3.88 | 4.88 | 0.64 | 0.87 |
| **Slap on the wrist** | **4.81** | **5.58** | **5.53** | **5.47** | **0.64** | **0.75** |
| Slip through the cracks | 4.86 | 5.14 | 5.24 | 5.19 | 0.68 | 0.31 |
| **Sound as a pound** | **3.93** | **4.91** | **2.51** | **2.50** | **0.20** | **0.20** |
| Spanner in the works | 4.53 | 5.49 | 4.84 | 4.24 | 0.32 | 1.00 |
| **Taken to the cleaners** | **4.33** | **5.00** | **5.40** | **3.00** | **0.12** | **0.00** |
| Throw in the towel | 5.42 | 6.02 | 5.69 | 4.18 | 0.32 | 0.63 |
| **Tip of the iceberg\*** | **5.21** | **6.26** | **5.79** | **5.18** |  |  |
| Up in the air | 4.70 | 5.38 | 4.84 | 3.65 | 0.28 | 0.29 |
| **Up sticks** | **2.86** | **4.40** | **2.56** | **2.35** | **0.12** | **0.00** |
| Up the duff | 5.28 | 5.95 | 2.69 | 2.31 | 0.16 | 0.50 |
| **Up the spout** | **2.86** | **3.50** | **3.50** | **1.88** | **0.08** | **0.00** |
| Water under the bridge | 4.98 | 6.00 | 5.67 | 3.59 | 0.24 | 0.50 |
| **Watering hole** | **4.14** | **4.47** | **5.00** | **4.41** | **0.56** | **0.15** |
| Wet behind the ears | 2.40 | 3.36 | 4.40 | 2.29 | 0.12 | 0.67 |
| **Who wears the trousers** | **5.27** | **6.12** | **5.51** | **4.06** | **0.44** | **0.60** |

\* Due to a printing error categorical decomposability ratings were not obtained for this item.

1. To alleviate concerns about using a parametric test, analyses were also conducted using a Kruskal-Wallis test for the likert data and a chi-square for the categorical data. The pattern of results was identical to those obtained using the parametric MANOVA. [↑](#footnote-ref-1)