

How often is ‘sometimes’?

Interpreting frequency expressions in language proficiency descriptors

Gerard Seinhorst

*Netherlands Defence Language Centre
Breda*

August 2026

ABSTRACT

Language proficiency descriptors commonly use frequency expressions such as *usually*, *generally*, *sometimes* and *typically* to describe expected language performance. Although such expressions are inherently approximate, relatively little empirical evidence is available on the extent to which users share similar interpretations of them. This study investigates how English frequency expressions used in language proficiency descriptors are interpreted by their users.

An online survey was conducted among 79 respondents representing a broad range of professional backgrounds and experience with language proficiency descriptors. Participants assigned numerical values to 20 frequency expressions presented within a carrier phrase and identified those they considered difficult to interpret.

The findings show a substantially greater degree of agreement than is often assumed. Agreement was greatest for the endpoint expressions *never* and *always*, but many other expressions also produced remarkably similar interpretations despite the absence of formal numerical definitions. Greater variation occurred primarily among expressions occupying the lower-middle ranges of the frequency continuum. Some expressions perceived as difficult to interpret nevertheless showed relatively high levels of agreement in their numerical interpretation.

Overall, the study provides empirical evidence that commonly used frequency expressions are interpreted with a considerable degree of shared understanding among language professionals. Rather than supporting the replacement of such expressions by approximate percentage ranges, the findings identify expressions that are associated with particularly high levels of agreement, highlight those warranting greater care in descriptor writing, and discuss the implications for the development and use of language proficiency descriptors in NATO STANAG 6001 and other descriptor-based language proficiency frameworks.■

CONTENTS

1. INTRODUCTION	6
2. SURVEY DESIGN AND METHODOLOGY	8
2.1 Survey design	8
2.2 Respondents	8
2.3 Questionnaire	9
2.4 Data analysis	9
3. RESULTS	11
3.1 Numerical interpretation of frequency expressions	11
3.2 Agreement among respondents	12
3.3 Reported interpretation difficulty	13
3.4 Professional background and experience	14
4. DISCUSSION	15
4.1 Shared understanding without numerical precision	15
4.2 Semantic clustering and the nature of approximation	15
4.3 Not all frequency expressions function alike	16
4.4 Perceived interpretation difficulty and observed agreement	17
4.5 The importance of context	17
4.6 Professional calibration reinforces shared interpretation	17
4.7 Frequency expressions and variation in assessment outcomes	18
4.8 Limitations of the study	18
5. CONCLUSIONS	19
5.1 Shared understanding without numerical precision	19
5.2 Functional vagueness as a characteristic of frequency expressions	19
5.3 Several frequency expressions form overlapping semantic clusters	19
5.4 Perceived interpretation difficulty does not predict disagreement	20
5.5 Professional experience appears more important than familiarity alone	20
5.6 Directions for future research	20
5.7 Overall conclusion	20
6. RECOMMENDATIONS	21
6.1 Continue to use frequency expressions in proficiency descriptors	21
6.2 Promote shared interpretation through calibration and standardisation	21
6.3 Support descriptor writers through practical guidance	21
6.4 Final reflection	21
Appendix A. <i>Survey questionnaire</i>	22
Appendix B. <i>Summary statistics for all frequency expressions</i>	25
Appendix C. <i>Supplementary quantitative analyses and respondent reflections</i>	26

1. INTRODUCTION

Frequency expressions such as *usually*, *often* and *sometimes* are widely used in language proficiency descriptors to describe expected language performance. In this report, the term *frequency expressions* refers to natural-language expressions indicating approximate frequency rather than explicit numerical values.

The present study examines such expressions as they are used in NATO STANAG 6001, the language proficiency standard used throughout NATO and by many partner nations. Like other descriptor-based language proficiency frameworks, STANAG 6001 makes extensive use of frequency expressions in its proficiency descriptors, for example:

- *usually* unable to sustain comprehension of longer texts;
- *generally* speaks in a way that is appropriate to the situation;
- may *occasionally* need to request repetition or rephrasing;
- can *typically* satisfy simple, predictable, personal and accommodation needs.

The subtle differences between such expressions may carry meaningful distinctions in expected performance and therefore contribute to the differentiation between proficiency levels.

Research on verbal probability expressions has shown that people may interpret natural-language terms such as *certain*, *possible* and *unlikely* differently.¹ This raises the question of whether frequency expressions used in language proficiency descriptors are similarly subject to variation in interpretation. In language assessment contexts, testers and raters have observed that certain frequency expressions appear vague or open to interpretation. It is therefore plausible that differences in how such expressions are understood could be a major source of recurring differences between raters when rating comparable language performances. Conversely, a high degree of shared interpretation would provide empirical support for the continued use of these expressions in descriptor-based frameworks.

However, relatively little empirical evidence is available on how frequency expressions in language descriptors are interpreted in practice, particularly regarding the degree of agreement among users. Previous research has mainly focused on variation in interpretation rather than on the extent to which users actually converge on broadly similar understandings.²

Against this background, the present study addresses the following research questions:

1. How do language professionals numerically interpret commonly used English frequency expressions as they occur in language proficiency descriptors?
2. Which frequency expressions show the greatest and least agreement among respondents?
3. To what extent do professional background and experience with language proficiency descriptors influence respondents' interpretations of the expressions?
4. Which frequency expressions do respondents themselves perceive as difficult to interpret?
5. To what extent do respondents' perceptions of interpretation difficulty correspond with the observed variation in their numerical ratings?

How often is “sometimes”?

If one language professional interprets *usually* as meaning “about 70% of the time” and another as “almost always”, are they really understanding the descriptor differently? Or does language allow sufficient flexibility for both interpretations to coexist?

¹ See, e.g.: Mandel, D.R., *Assessment and Communication of Uncertainty in Intelligence to Support Decision-Making* (2020)

² See, e.g.: Bocklisch, F., Bocklisch, S.F. & Krems, J.F., Sometimes, often, and always: Exploring the vague meanings of frequency expressions. *Behav Res* 44, 144–157 (2012)

These questions were investigated through an online survey among language professionals mainly involved in language teaching and assessment. Participants assigned numerical values to the frequency expressions and were invited to comment on any expressions they found difficult to interpret.

The findings provide empirical evidence on the degree of shared interpretation of frequency expressions in language proficiency descriptors. They also have practical implications for descriptor development, standardisation practices and future revisions of NATO STANAG 6001 and other descriptor-based language proficiency frameworks.

2. SURVEY DESIGN AND METHODOLOGY

2.1 Survey design

An anonymous online survey was conducted among language professionals from the international language-testing and teaching community. Recruitment took place primarily through the *Bureau for International Language Coordination* (BILC) STANAG 6001 Testing Workshop, the BILC Conference, and related professional networks. BILC is NATO's coordinating body for cooperation on language education, testing and policy. The survey focused on 20 English frequency expressions commonly encountered in language proficiency descriptors (e.g. *often*, *sometimes* and *usually*).

The questionnaire combined quantitative and qualitative methods. Participants first completed a background questionnaire, then assigned numerical values to the frequency expressions, and finally identified expressions they considered difficult to interpret. The full questionnaire is reproduced in [Appendix A](#).

2.2 Respondents

A total of 79 respondents completed the survey. Participants represented a broad range of professional backgrounds, including language testers and raters, assessment specialists, language teachers, and other language professionals. Personal information was provided voluntarily; among those who supplied these details, 20 different nationalities were represented.

Respondents also differed considerably in their familiarity with language proficiency descriptors and in their professional experience in language teaching, testing, rating and descriptor development.

Respondent profile

[Figure 1](#) summarises the composition of the respondent sample. Although the survey was not intended to provide a statistically representative sample, it includes respondents with a broad range of professional backgrounds and experience.

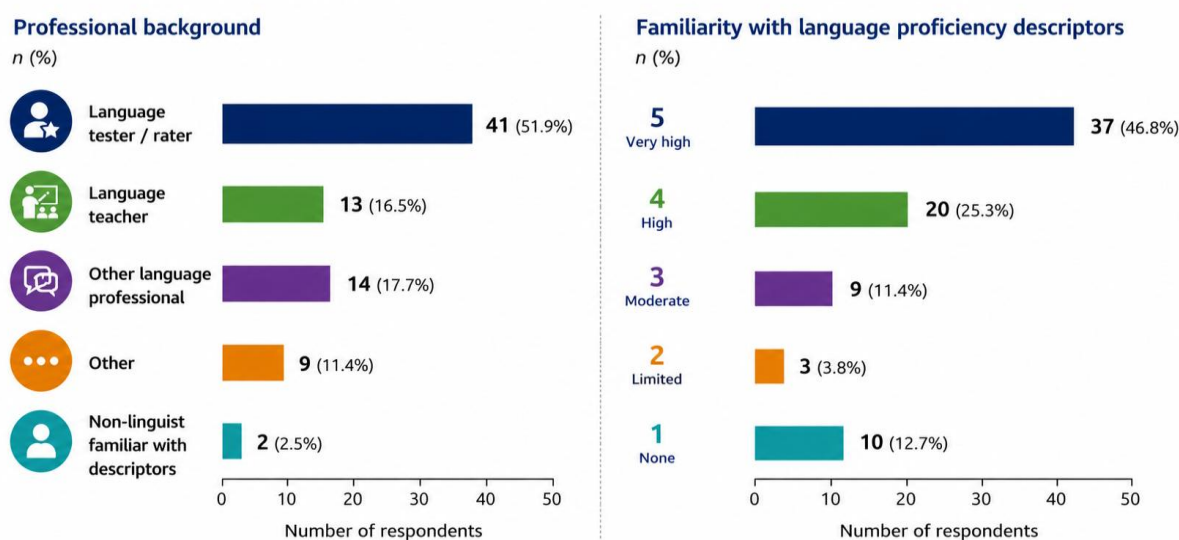


Figure 1: Respondent profile (N=79)

2.3 Questionnaire

The questionnaire consisted of three parts.

Part 1 – Background information

Participants first provided general background information, including their professional role, familiarity with language proficiency descriptors and other contextual information. This information was collected to explore whether professional background or familiarity with language proficiency descriptors influenced respondents' interpretations.

Part 2 – Numerical interpretation of frequency expressions

The core of the survey consisted of 20 English frequency expressions commonly encountered in descriptors. For each expression, participants were asked to assign a numerical value on a scale from 0 ("never") to 100 ("always"), representing the frequency they associated with that expression.

Rather than embedded in complete proficiency descriptors, the expressions were presented individually within a short carrier phrase: "*The language user can _____ complete the task.*" This allowed respondents' intuitive interpretations to be compared while minimising contextual influences. Some expressions fit this construction more naturally than others. For example, expressions such as *regularly* and *consistently* often describe repeated or sustained behaviour rather than the frequency of a single event. This limitation was accepted because maintaining an identical context for all expressions was considered more important than optimising the wording for individual items.

The numerical ratings were not intended to establish numerical definitions. Instead, they provided a common measurement scale for comparing respondents' interpretations and quantifying agreement.

Part 3 – Qualitative comments

Finally, respondents were asked to indicate which expressions they considered difficult to interpret and to explain their reasoning in their own words. These comments provided valuable additional insight into the quantitative findings and helped identify recurring themes that could not easily be captured through numerical analysis alone.

2.4 Data analysis

The survey generated two complementary types of data: numerical ratings and qualitative comments. Both were analysed to provide a more comprehensive understanding of the interpretation of frequency expressions.

As a quality-control measure, responses were screened for internal consistency by examining whether individual ratings followed the expected ordinal relationship between clearly ordered frequency expressions (e.g. *never* < *rarely* < *sometimes* < *often* < *always*). The data showed a high degree of internal consistency, with only a small number of isolated reversals that were generally minor and did not indicate systematic misunderstanding of the task. Consequently, all responses were retained for analysis.

Descriptive statistics

For each expression, the mean, median, mode, standard deviation (SD) and interquartile range (IQR) were calculated. The *median* was used as the principal measure of central tendency, as it is robust to unusually high or low responses. Since the study focused primarily on agreement between respondents, the IQR was adopted as the principal measure of dispersion.

Unlike the standard deviation, it is less affected by isolated extreme values.

Why the IQR is a useful measure

Imagine that almost all respondents interpret an expression as meaning between 70 and 80, while one respondent assigns a value of 20. The standard deviation increases noticeably because of this single extreme rating. The IQR, however, changes very little because it reflects only the middle half of all responses. Since the purpose of this study is to examine the degree of shared interpretation rather than the influence of isolated responses, the IQR provides a particularly informative measure of agreement.

Although a small number of ratings lay outside the main response patterns, these had little influence on the reported results. All responses were therefore retained in the analyses.

Exploratory analyses

Supplementary analyses examined whether interpretation varied according to respondents' professional background, familiarity with language proficiency descriptors, and their use of English as a first or foreign language (see [Appendix C](#)). Qualitative comments were analysed thematically to identify recurring observations about the interpretation and use of the expressions. As these analyses were exploratory rather than hypothesis-driven, they are interpreted with appropriate caution.

3. RESULTS

This chapter presents the principal findings of the survey. The results are organised according to the research questions. First, the overall numerical interpretation of the frequency expressions is presented, followed by analyses of respondent agreement, self-reported interpretation difficulty and the influence of professional background. The broader interpretation of these findings and their implications are discussed in [Chapter 4](#).

3.1 Numerical interpretation of frequency expressions

The primary objective of the survey was to investigate how language professionals interpret frequency expressions commonly used in language proficiency descriptors.

[Figure 2](#) presents the complete distributions of numerical ratings for all 20 expressions in the form of a *ridgeline plot*. Unlike summary tables or bar charts, the ridgeline plot simultaneously displays the typical interpretation of each expression, the degree of agreement among respondents and the overall shape of the response distribution. Narrow ridges indicate relatively strong agreement, whereas broader or multi-peaked distributions indicate greater diversity in interpretation. Expressions are ordered by their median values. The white dot represents the median, the black horizontal bar the IQR, and the value shown to the right of each ridge is the mean. Detailed summary statistics are provided in [Appendix B](#).

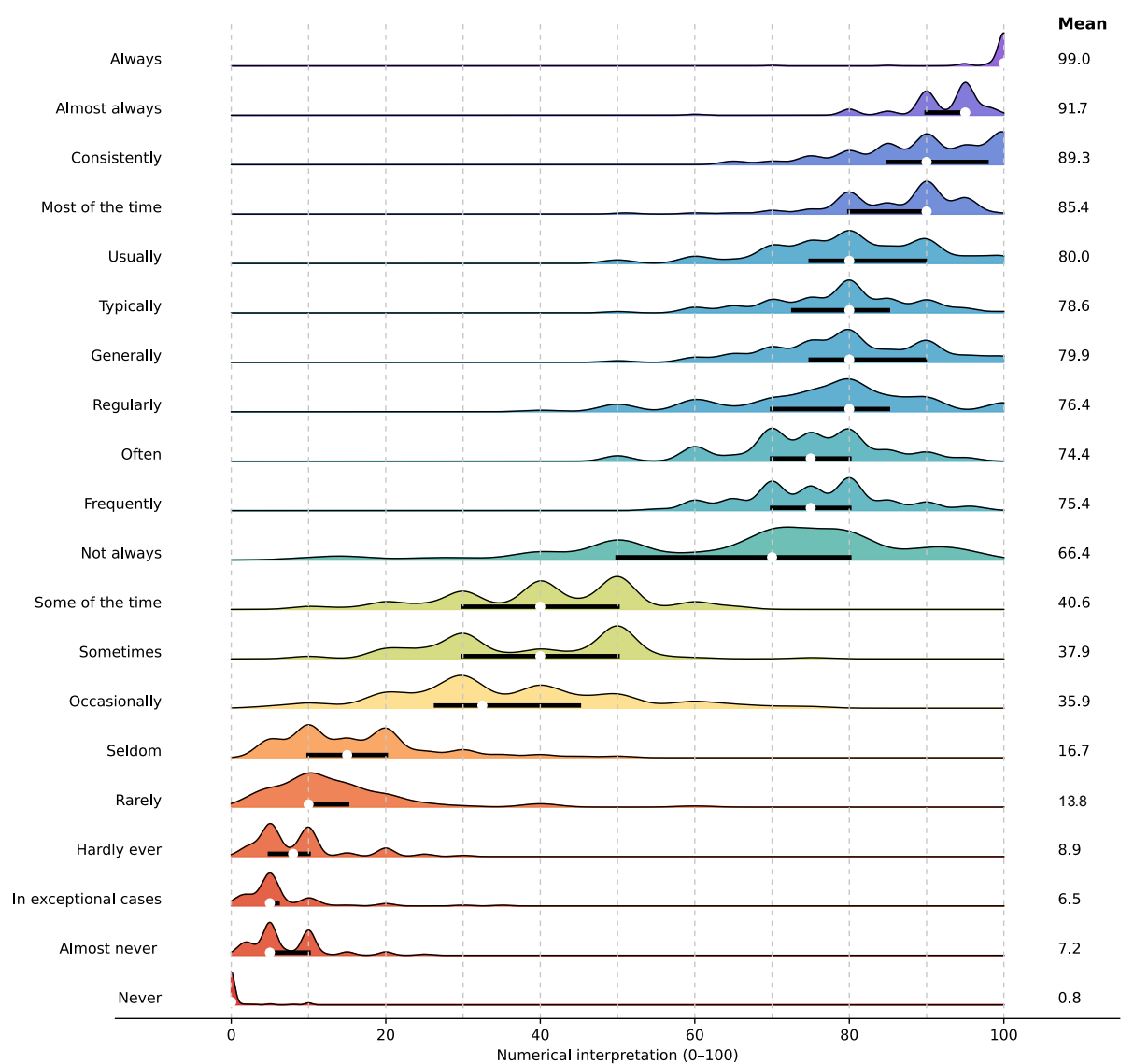


Figure 2: Numerical interpretation of frequency expressions (0-100)

Several general observations emerge from the figure.

First, nearly all expressions exhibit clear central tendencies despite individual variation in the numerical ratings.

Second, the endpoint expressions *never* and *always* produced the greatest agreement. Although most respondents assigned values of 0 and 100 respectively, a small number assigned values slightly above 0 for *never* or slightly below 100 for *always*.

Third, the greatest variation occurred among expressions occupying the lower middle of the frequency continuum. These distributions were generally broader than those near the endpoints and, in several cases, showed more than one local peak.

Another notable feature of the distributions is that virtually none of the expressions were centred around the midpoint of the numerical continuum. Instead, the expressions clustered in lower or higher frequency regions, with few appearing to represent frequencies close to “about half of the time”.

In addition, analysis of the full dataset revealed a pattern concerning the numerical precision of responses. Intermediate expressions were predominantly assigned rounded values such as 40, 50, 60 and 70, whereas expressions near the endpoints (e.g. *almost never*, *in exceptional cases* and *almost always*) more frequently received precise values such as 1, 3, 5, 95, 97 and 99. Consistent with this pattern, the modal value for every expression was a multiple of ten except for *almost never* (5), *in exceptional cases* (5) and *almost always* (95).

3.2 Agreement among respondents

Figure 3 complements the ridgeline plot by ranking the expressions according to their interquartile range, thereby facilitating comparison of respondent agreement.

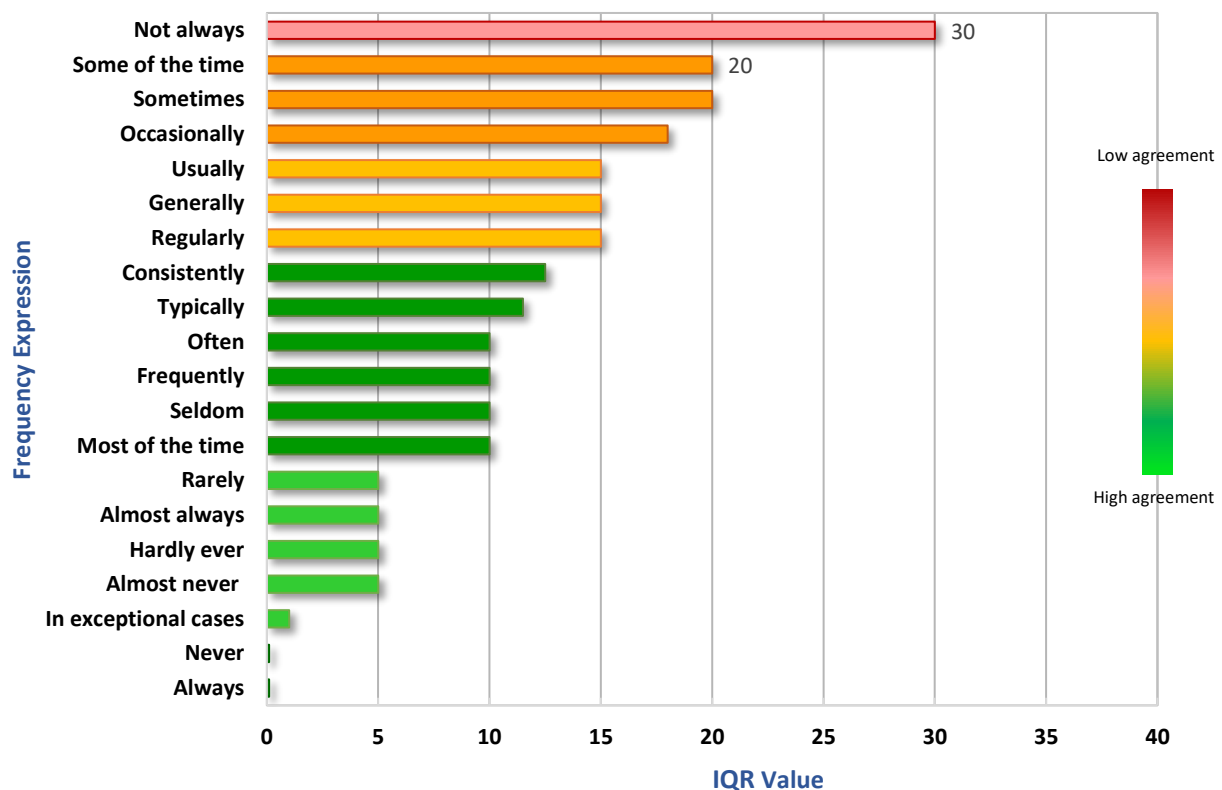


Figure 3: Degree of agreement among respondents in the interpretation of the frequency expressions

As expected, *never* and *always* exhibited the smallest interquartile ranges, indicating the greatest agreement among respondents. Several expressions located towards the middle of the frequency continuum showed broader IQRs, reflecting greater variation in numerical interpretation.

Even the expressions with the largest IQRs remained concentrated within recognisable regions of the frequency continuum rather than being distributed evenly across the entire scale.

3.3 Reported interpretation difficulty

After completing the numerical rating task, respondents were asked to indicate which, if any, of the frequency expressions they had found difficult to interpret and to explain their choices.

Figure 4 plots each expression according to two dimensions: the number of respondents who identified it as difficult to interpret and the observed IQR of its numerical ratings. Expressions positioned higher on the vertical axis were more frequently perceived as difficult to interpret, whereas expressions further to the right showed greater variation in numerical interpretation.

The expressions most commonly identified as difficult to interpret were *typically*, *some of the time*, *in exceptional cases*, *occasionally*, *consistently* and *not always*. By contrast, expressions such as *never*, *always* and *almost always* were mentioned by relatively few respondents as difficult to interpret. Some expressions that respondents described as difficult, such as *typically*, nevertheless showed relatively narrow interquartile ranges, whereas other expressions with similar response distributions, such as *usually*, attracted comparatively few comments.

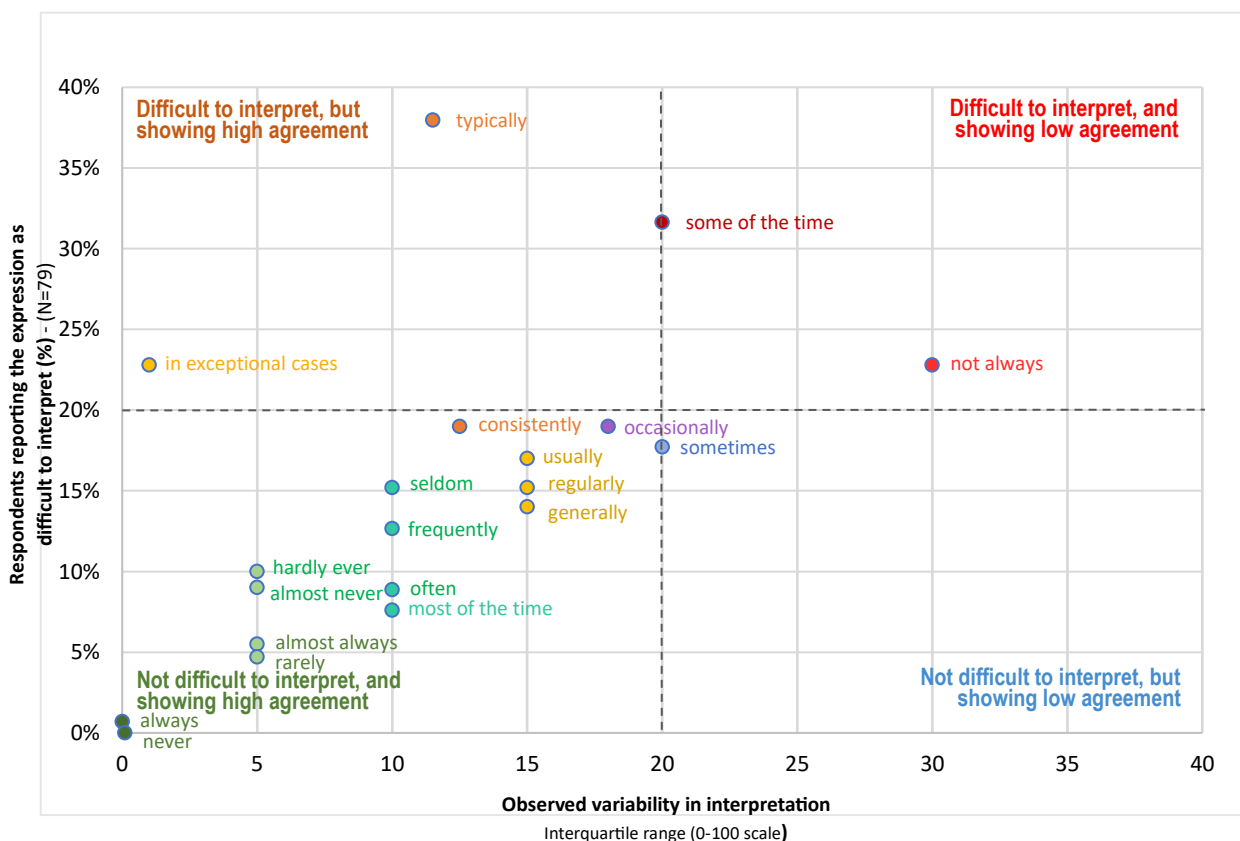


Figure 4: Reported interpretation difficulty and observed agreement for the 20 frequency expressions.

3.4 Professional background and experience

Additional analyses explored whether respondents' professional background or their familiarity with language proficiency descriptors influenced the consistency of their interpretations (see [Appendix C](#)).

Respondents from different professional groups showed broadly similar patterns of interpretation. Respondents working as language testers or raters exhibited slightly greater agreement with one another than teachers and respondents from other professional backgrounds, particularly for expressions occupying the middle of the frequency continuum. By contrast, self-reported familiarity with proficiency scales showed only modest associations with rating consistency.

Similarly, respondents who use English as their first language and those who use English as a foreign language produced highly comparable interpretations. No systematic differences were identified that would suggest fundamentally different interpretations of the frequency expressions.

Overall, differences between respondent groups were modest.

4. DISCUSSION

The purpose of this study was not to assign fixed numerical meanings to frequency expressions, but to investigate the extent to which language professionals share common interpretations of the expressions used in language proficiency descriptors. The results provide a basis for considering what the observed variation and agreement mean for the role of frequency expressions in descriptor-based language assessment. In particular, they raise questions about how approximate language can support shared professional understanding without relying on fixed numerical definitions.

4.1 Shared understanding without numerical precision

The principal finding of the study is that language professionals show a substantial degree of shared understanding in their interpretation of frequency expressions. Although respondents were required to assign precise numerical values, the resulting distributions showed clear central tendencies for almost all expressions and relatively modest variation.

This distinction between numerical precision and shared understanding is central to the interpretation of the findings. Frequency expressions do not require fixed percentage definitions in order to communicate effectively. Instead, respondents appear to possess stable intuitions about the approximate frequencies represented by many commonly used expressions. For descriptor-based assessment, this shared understanding is arguably more important than numerical precision, because descriptors are intended to support informed professional judgement rather than prescribe mathematically exact thresholds.

One possible explanation for this shared understanding is that many of the expressions investigated belong to the core vocabulary of English and are encountered routinely in everyday communication. Shared intuitions about their meanings are therefore likely to develop long before language professionals encounter them in proficiency descriptors. Professional experience with descriptors may reinforce this shared understanding further, but the findings do not suggest that such experience is necessary for it to develop.

4.2 Semantic clustering and the nature of approximation

Taken together, the numerical distributions reveal that many frequency expressions occupy overlapping regions of the frequency continuum rather than isolated numerical positions (Figure 5). Several groups of expressions were interpreted so similarly that respondents themselves frequently commented on the difficulty of distinguishing between them.

High frequency expressions including *usually*, *frequently*, *often* and *typically* formed one closely related group. Likewise, *seldom*, *rarely*, *hardly ever*, *in exceptional cases* and *almost never* clustered closely together at the lower end of the continuum. Rather than representing sharply separated concepts, these expressions appear to function as overlapping semantic alternatives.

The survey design deliberately required respondents to assign a separate numerical value to every expression. In ordinary language use, however, such expressions are rarely interpreted as representing distinct numerical categories. Instead, speakers appear to treat them as belonging to broader semantic regions whose boundaries overlap naturally.

This finding helps explain why several expressions were reported as difficult to distinguish despite producing remarkably similar numerical ratings. The perceived difficulty therefore seems to reflect the challenge of discriminating between near-synonyms rather than uncertainty about their overall meaning.

More generally, these findings indicate that approximation and synonymy are functional characteristics of natural language rather than shortcomings of descriptor language. If descriptor writers intended to specify exact percentages, they could simply state numerical values directly. The deliberate use of frequency expressions serves a different communicative purpose: describing approximate patterns of performance while allowing sufficient flexibility to accommodate authentic language use.

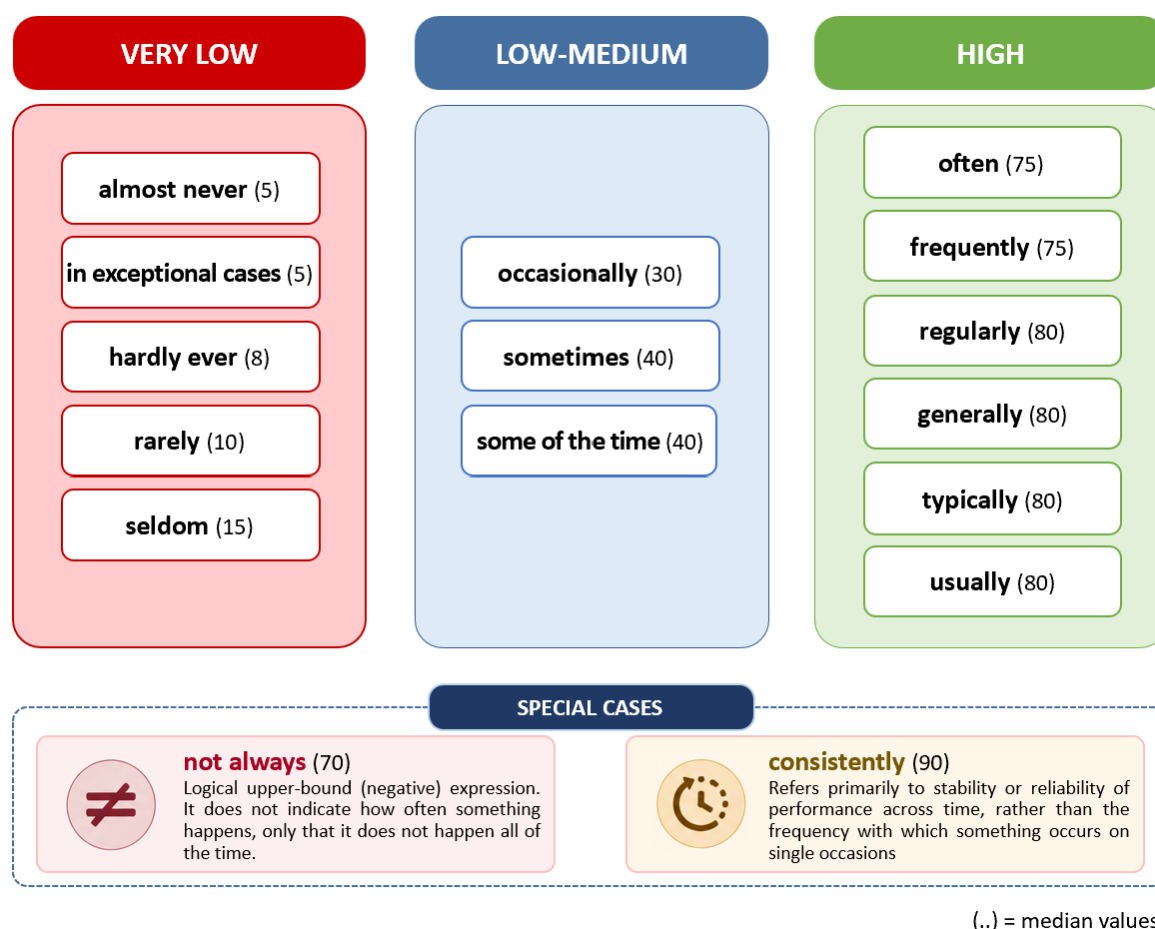


Figure 5: Semantic groupings of frequency expressions with similar median ratings.

4.3 Not all frequency expressions function alike

Although most expressions followed the general pattern described above, two expressions stood apart from the remainder.

The first is *not always*. Unlike conventional frequency expressions, *not always* does not directly indicate how often something occurs. Logically, it merely excludes one possibility – that something occurs in every case. This allows interpretations ranging from *almost always* to frequencies well below 50, with some ratings as low as 10 or 15, which was reflected in the unusually broad distribution of responses. Respondents also identified this expression comparatively often as difficult to interpret.

Despite this structural difference, *not always* nevertheless appears to function in practice as a conventional frequency expression. With a median value of 70, *not always* occupied a similar region of the frequency continuum as some higher-frequency expressions. However, its broad distribution revealed substantial variation in the range of interpretations associated with it.

A second exception is *consistently*. Several respondents observed that this expression differs fundamentally from the other items because it refers primarily to the stability or reliability of performance rather than to the frequency with which a behaviour occurs. Assigning a single percentage therefore felt less intuitive for many respondents. Within language proficiency descriptors, *consistently* may therefore be interpreted less as a frequency expression than as an indication of sustained or dependable performance.

Interestingly, respondents commented only on *consistently*, despite the fact that other expressions also extend beyond simple frequency. For example, *regularly* likewise refers primarily to recurring behaviour rather than to the frequency of a single event, yet it attracted no comparable observations.

One possible explanation is that *regularly* has become more strongly associated with frequency in everyday language, whereas *consistently* continues to evoke notions of uniformity or reliability. Although the present study was not designed to investigate this distinction directly, the findings suggest that respondents accepted *regularly* as functioning effectively as a frequency expression despite its somewhat different primary meaning.

These findings illustrate that expressions occupying similar positions on the numerical continuum may nevertheless differ in the way respondents conceptualise them.

4.4 Perceived interpretation difficulty and observed agreement

One of the more interesting findings is that respondents' subjective impressions did not always correspond with the consistency of their numerical interpretations. Several expressions, including *typically*, *consistently* and *some of the time*, were frequently described as difficult to interpret, yet they nevertheless produced relatively similar numerical ratings. Conversely, expressions such as *usually* and *sometimes* exhibited broader numerical distributions while attracting comparatively few comments.

This apparent discrepancy suggests that perceived interpretation difficulty and observed agreement represent different aspects of language processing. Respondents may recognise that certain expressions overlap semantically and therefore find it difficult to distinguish between them, even though they nevertheless assign broadly similar meanings.

Similarly, broader interquartile ranges should not automatically be interpreted as evidence of problematic vagueness. Even the expressions showing the greatest variation occupied well-defined regions of the frequency continuum. In most cases, respondents appeared to agree on the general location of an expression while differing mainly in the breadth of the range they associated with it. In other words, disagreement typically concerned the boundaries of an expression rather than its overall position on the continuum.

4.5 The importance of context

Many respondents emphasised that frequency expressions cannot be interpreted independently of the descriptors in which they occur. Their meaning is influenced by factors such as the communicative task, the linguistic feature being described, and the proficiency level concerned. This observation is consistent with descriptor-based approaches to language assessment, which evaluate language behaviour within authentic communicative contexts rather than in isolation.

At the same time, one of the more noteworthy findings of the present study is that respondents demonstrated considerable agreement even though contextual information had been deliberately removed. This suggests that commonly used frequency expressions possess relatively stable lexical meanings that remain recognisable outside individual descriptors. Context undoubtedly refines interpretation, but the underlying semantic understanding appears to be shared to a considerable extent before contextual information is taken into account.

4.6 Professional calibration reinforces shared interpretation

The analyses indicate that practical experience as a language tester or rater was associated with slightly greater consistency of interpretation than self-reported familiarity with language proficiency frameworks. Although the differences were modest and no causal relationship can be established, the overall pattern is consistent with the view that repeated application of descriptors gradually reinforces a shared operational understanding of their language.

This interpretation is supported by observations frequently made during moderation and standardisation sessions. Experienced raters occasionally discover that the wording of a descriptor differs from how they remember it, despite having applied it successfully for many years. Such

experiences indicate that professionals internalise the intended meaning of descriptors rather than their precise wording.

For organisations such as BILC, where language professionals from diverse linguistic and cultural backgrounds work within common assessment frameworks, this is an encouraging finding. It indicates that shared professional practice may contribute as much to shared interpretation as the wording of the descriptors themselves.

4.7 Frequency expressions and variation in assessment outcomes

The findings also have implications for a possible source of variation in descriptor-based assessment. When raters assign different proficiency levels to comparable language performances, one possible explanation is that they interpret the frequency expressions in the relevant descriptors differently. The present study was not designed to investigate rater agreement directly and therefore cannot establish whether variation in the interpretation of frequency expressions leads to variation in assessment outcomes.

Nevertheless, the substantial degree of agreement observed for most expressions indicates that divergent interpretations of individual frequency expressions are unlikely, by themselves, to be a major source of variation between raters' proficiency-level decisions for comparable language performances. This does not mean that such variation cannot occur, nor that frequency expressions are irrelevant to assessment practice. Variation in assessment outcomes may also reflect a combination of other factors, including the assessment context and task, the interpretation and application of the descriptor as a whole, and organisational or institutional considerations that may influence the proficiency level awarded in addition to the assessment evidence.

The results therefore provide no clear reason to regard the approximate nature of frequency expressions as a major obstacle to reliable descriptor-based assessment. Rather than seeking to eliminate this approximation through numerical definitions, efforts to support shared assessment practice are likely to benefit from attention to the broader conditions in which descriptors are interpreted and applied.

4.8 Limitations of the study

Several limitations should be considered when interpreting the findings. First, respondents were asked to assign numerical values to inherently approximate expressions. Although this provided a common measurement scale for comparing interpretations, it necessarily simplified the way such expressions are normally understood in authentic communication.

Secondly, the expressions were presented in isolation rather than embedded within complete language proficiency descriptors. This deliberate design choice enabled the study to investigate their underlying lexical interpretation while minimising contextual influences. However, respondents themselves frequently emphasised that context contributes to the interpretation of frequency expressions in authentic assessment settings.

Thirdly, participation was voluntary. Although the respondents represented a broad range of professional backgrounds and nationalities, volunteers may have had a greater interest in descriptor language than the wider population of potential users, which may have affected the representativeness of the sample.

Finally, while the number of respondents was sufficient to identify general patterns of interpretation, the sample was not intended to be statistically representative of the wider population of language professionals. The findings should therefore be interpreted as evidence of shared tendencies within this professional community rather than as definitive estimates of how all language users would interpret these expressions.

5. CONCLUSIONS

This study investigated how language professionals interpret frequency expressions commonly used in language proficiency descriptors. Rather than attempting to assign precise numerical equivalents, the study examined the extent to which users of proficiency descriptors share common interpretations of their meanings.

The findings lead to the following overarching conclusions.

5.1 Shared understanding without numerical precision

The principal conclusion of this study is that language professionals show a substantial degree of shared understanding in their interpretation of frequency expressions. Although respondents were required to assign precise numerical values, the resulting distributions showed clear central tendencies for almost all expressions and relatively modest variation.

The findings indicate that frequency expressions do not require fixed numerical definitions in order to convey meaning effectively. Respondents generally interpreted the expressions as representing approximate regions of the frequency continuum rather than exact percentages. The practical value of these expressions therefore lies not in their ability to specify precise numerical thresholds, but in their ability to evoke broadly similar interpretations among users.

For descriptor-based assessment, this shared understanding is arguably more important than numerical precision, because descriptors are intended to support informed professional judgement rather than prescribe mathematically exact criteria.

5.2 Functional vagueness as a characteristic of frequency expressions

The findings also challenge the assumption that linguistic vagueness necessarily reduces clarity or effectiveness. Frequency expressions describe typical patterns of behaviour rather than fixed quantities. Their purpose is not to identify precise points on a numerical scale, but to indicate approximate regions of frequency that allow users to interpret language performance in a shared way. The observed distributions further illustrate this characteristic: the expressions did not map evenly onto the numerical continuum, but instead clustered around particular regions of frequency. This suggests that frequency expressions function as meaningful linguistic categories rather than as labels for predefined numerical intervals. Their flexibility enables descriptors to accommodate the variation that naturally occurs in authentic language use, while still providing users with a shared basis for interpretation. Attempts to replace such expressions with approximate percentage ranges or numerical brackets may therefore create an impression of objectivity that is inconsistent with the complex and context-dependent nature of language performance. The effectiveness of frequency expressions depends not on eliminating their inherent approximation, but on maintaining the shared professional understanding that allows them to function effectively.

5.3 Several frequency expressions form overlapping semantic clusters

The survey identified several clearly recognisable groups of expressions that were interpreted very similarly (summarised in [Figure 5](#)). Expressions such as *usually*, *generally*, and *typically* tend to be interpreted in broadly similar ways, as do *hardly ever*, *rarely* and *almost never*. These clusters were reflected both in the numerical data and in respondents' qualitative comments, many of whom reported difficulty distinguishing between near-synonymous expressions. Rather than representing a weakness of descriptor language, this overlap appears to provide useful linguistic flexibility, allowing descriptor writers to vary their wording without substantially altering the intended meaning.

Two expressions nevertheless stood apart from the remainder. Both *not always* and *consistently* attracted a disproportionate number of qualitative comments, and their numerical response patterns also differed from those of most other expressions. Taken together, these findings suggest that respondents perceived these two expressions somewhat differently from the remainder of the frequency expressions investigated.

The broad distribution observed for *not always*, together with respondents' comments, suggests that many participants interpreted it less as a conventional frequency expression than as a negated expression defining only the upper boundary of the continuum. Similarly, the repeated comments concerning *consistently* suggest that many respondents associated it primarily with stability or reliability across repeated performances rather than with frequency itself.

5.4 Perceived interpretation difficulty does not predict disagreement

A further conclusion is that respondents' subjective impressions of interpretation difficulty did not necessarily correspond with observed agreement. Several expressions that respondents frequently described as difficult to interpret nevertheless produced relatively similar numerical interpretations, whereas some expressions showing greater variation attracted comparatively few comments.

Subjective perceptions of vagueness should therefore not be taken as reliable indicators of actual disagreement. Users' own impressions and quantitative evidence appear to capture different aspects of how frequency expressions are understood.

5.5 Professional experience appears more important than familiarity alone

Differences associated with self-reported familiarity with proficiency scales were relatively modest. More noticeable patterns were observed among respondents with practical experience as language testers or raters, who generally showed slightly greater agreement with one another when interpreting the expressions.

Although no causal relationship can be established, this pattern is consistent with the view that repeated engagement with descriptor-based assessment, moderation and standardisation practices may contribute to the development and reinforcement of shared interpretations of descriptor language. Familiarity with a proficiency framework alone, by contrast, appeared to be less closely associated with agreement among respondents.

The findings therefore suggest that shared interpretation depends less on knowledge of a proficiency framework alone than on continued professional engagement with the practical use of its descriptors.

5.6 Directions for future research

The present study represents an initial exploration of how frequency expressions are interpreted in isolation. Further research could extend these findings by examining how such expressions are interpreted within complete language proficiency descriptors and authentic assessment contexts.

Future studies might also investigate how shared interpretations develop through professional experience, calibration and standardisation, whether comparable patterns occur in other languages, and how other types of qualitative descriptor language – such as degree, quantity or probability expressions – are interpreted by users of proficiency frameworks.

Such work would contribute to a broader empirical understanding of the language on which descriptor-based assessment depends.

5.7 Overall conclusion

Taken together, the findings indicate that frequency expressions support a meaningful degree of shared interpretation without requiring fixed numerical definitions. Their effectiveness therefore depends less on numerical specification than on the shared professional understanding developed through experience, practice and discussion. This central conclusion provides the basis for the recommendations presented in [Chapter 6](#).

6. RECOMMENDATIONS

The findings of this study lead to several practical recommendations for the development, interpretation and use of language proficiency descriptors. Although the survey focused on frequency expressions in NATO STANAG 6001, the recommendations are equally relevant to other descriptor-based language proficiency frameworks that rely on frequency expressions to describe expected performance.

6.1 Continue to use frequency expressions in proficiency descriptors

The findings provide no evidence that commonly used frequency expressions should be replaced by approximate percentage ranges or fixed numerical definitions. Although some respondents expressed a wish for more explicit numerical guidance when interpreting and applying frequency expressions in descriptor-based assessment, the results indicate that such guidance would be unlikely to improve the practical use of proficiency descriptors.

Frequency expressions are inherently approximate. Their purpose is not to communicate mathematically precise values but to describe patterns of performance that naturally vary across tasks, contexts and individuals. Consequently, assigning numerical equivalents to individual expressions would risk creating an impression of objectivity that is inconsistent with the complex and context-dependent nature of language performance.

Descriptor developers should therefore retain frequency expressions where they provide clear and meaningful descriptions of language performance, rather than replacing them with numerical definitions or percentage ranges.

6.2 Promote shared interpretation through calibration and standardisation

The study provides some evidence that shared interpretation depends less on numerical specification than on shared professional understanding developed through experience and professional practice. Repeated experience in applying descriptors may reinforce this understanding. Organisations should therefore continue to invest in rater training, moderation and standardisation activities that provide opportunities for discussion, calibration and the development of shared professional practices.

6.3 Support descriptor writers through practical guidance

Although the survey does not support prescribing numerical meanings for individual expressions, it does identify patterns that may assist descriptor writers. These patterns can be translated into practical guidance to support more consistent and informed choices when developing descriptors. Such guidance could include examples of expressions that are generally interpreted as near-synonyms and may therefore be used interchangeably where stylistically appropriate. It could also provide advice on selecting frequency expressions associated with greater agreement among users, as well as notes on expressions that warrant particular care because they attracted broader or more varied interpretations. Particular attention may be given to *negated expressions*, such as *not always*, whose logical structure differs from that of conventional frequency expressions and which may therefore be interpreted in a wider range of ways.

6.4 Final reflection

Frequency expressions are sometimes criticised for lacking the precision expected of formal assessment language. The findings of this study suggest, however, that their approximate nature should not be viewed as a deficiency to be eliminated. Their inherent vagueness is not a weakness but a functional characteristic that allows proficiency descriptors to represent naturally variable language performance without imposing artificial precision.

The challenge for descriptor-based assessment is therefore not to eliminate this functional vagueness, but to preserve the shared professional understanding that allows it to function effectively.

Appendix A. *Survey questionnaire*

Interpreting Frequency Expressions in Proficiency Descriptors

This survey investigates how people interpret frequency expressions used in language proficiency descriptors (e.g. *often, typically, rarely, most of the time*).

For each expression, participants are asked to estimate how many cases out of 100 they personally think the expression refers to.

The survey is intended to support clearer and more consistent interpretation of descriptors in language teaching, testing, and assessment contexts.

The questionnaire takes approximately 10 minutes.

Participation is voluntary. Responses will be analyzed and reported in anonymized form. You may optionally provide your name or email address for follow-up communication or to receive information about the study results. Any identifying information will not be included in reports or publications.

By continuing with the questionnaire, you consent to the use of your responses for research, analysis, and reporting purposes.

*** Mandatory question**

1. **Your name (optional):** _____

2. **Your e-mail address (optional):** _____

3. **Your background ***

Select one

- ☐ Language tester / rater
- ☐ Language teacher
- ☐ Other language professional
- ☐ Non-linguist familiar with language proficiency scales
- ☐ Other

4. **Experience**

How many years of experience do you have in language teaching, testing or assessment?

Select one

- ☐ None
- ☐ 1-3 years
- ☐ 4-10 years
- ☐ 11-20 years
- ☐ More than 20 years

5. **Familiarity with language descriptors**

How familiar are you with language proficiency descriptors/scales (e.g., STANAG 6001, CEFR, ACTFL)

Select one

	1	2	3	4	5	
Not at all	☐	☐	☐	☐	☐	Very familiar

INSTRUCTIONS

Below are 20 English frequency expressions presented in random order. For each expression, indicate approximately how many cases out of 100 you think the expression refers to.

For example, consider the expression 'occasionally' in the sentence: *The language user can occasionally complete the task.*

Answer with a *whole number* between 0 (= never) and 100 (= always). For instance, if you think that 'occasionally' means that something happens in 45 out of 100 cases, your answer would be 45.

There are no correct or incorrect answers; the study is interested in your personal interpretation of the expressions.

If you would like to add any comments, you may do so at the end of the questionnaire.

For each frequency expression given below, imagine the following sentence with that expression filling the blank: *The language user can _____ complete the task.*

Out of 100 cases, how many cases do you think this expression refers to?

1. **Sometimes**
 2. **Almost never**
 3. **Regularly**
 4. **Not always**
 5. **Hardly ever**
 6. **Most of the time**
 7. **Seldom**
 8. **Generally**
 9. **In exceptional cases**
 10. **Frequently**
 11. **Often**
 12. **Always**
 13. **Some of the time**
 14. **Consistently**
 15. **Occasionally**
 16. **Almost always**
 17. **Rarely**
 18. **Typically**
 19. **Never**
 20. **Usually**
-

ADDITIONAL QUESTIONS

6. **Which expressions did you find difficult to interpret? ***

(more than one answer possible)

- ☐ sometimes
- ☐ almost never
- ☐ regularly
- ☐ not always
- ☐ hardly ever
- ☐ most of the time
- ☐ seldom
- ☐ generally
- ☐ in exceptional cases
- ☐ frequently
- ☐ often
- ☐ some of the time
- ☐ consistently
- ☐ occasionally
- ☐ almost always
- ☐ rarely
- ☐ typically
- ☐ usually
- ☐ none of the above

7. **Please explain why you found these expression(s) difficult to interpret.**

COMMENTS

8. Please add any comments you may have on the use of frequency expressions in language proficiency descriptors.

THIS IS THE END OF THE QUESTIONNAIRE

Thank you very much for taking the time to complete this questionnaire.

Your participation is greatly appreciated and will contribute to a clearer and more consistent descriptor interpretation among language professionals and raters.

If you provided your email address, you may be contacted with a brief summary of the findings once the study has been completed.

Thank you again for your valuable contribution.

Appendix B. *Summary statistics for all frequency expressions*

A total of 79 respondents assigned numerical values (0–100) to each of the 20 frequency expressions included in the survey. Throughout this report, the median and interquartile range (IQR) are used as the principal measures of central tendency and dispersion because they are less affected by extreme responses than the mean and standard deviation, and better reflect the interpretation of approximate linguistic expressions. The complete summary statistics are provided below for reference.

Rank	Frequency expression	Median	Mean	Mode	IQR	SD	Min	Max
1	<i>Never</i>	0	0.8	0	0.0	2.5	0	10
2	<i>In exceptional cases</i>	5	6.5	5	1.0	5.8	0	35
3	<i>Almost never</i>	5	7.2	5	5.0	4.8	1	25
4	<i>Hardly ever</i>	8	8.9	10	5.0	6.0	1	30
5	<i>Rarely</i>	10	13.8	10	5.0	9.7	2	50
6	<i>Seldom</i>	15	16.7	10	10.0	9.9	5	50
7	<i>Occasionally</i>	30	35.9	30	18.0	14.5	5	75
8	<i>Sometimes</i>	40	37.9	50	20.0	12.7	10	75
9	<i>Some of the time</i>	40	40.6	50	20.0	12.5	10	65
10	<i>Not always</i>	70	66.4	70	30.0	19.3	10	95
11	<i>Often</i>	75	74.4	70	10.0	9.7	50	95
12	<i>Frequently</i>	75	75.4	80	10.0	9.1	55	97
13	<i>Regularly</i>	80	76.4	80	15.0	13.2	40	100
14	<i>Typically</i>	80	78.6	80	12.3	9.6	50	100
15	<i>Generally</i>	80	79.9	80	14.8	10.4	50	100
16	<i>Usually</i>	80	80.0	80	14.8	10.9	50	100
17	<i>Most of the time</i>	90	85.4	90	10.0	8.7	51	99
18	<i>Consistently</i>	90	89.3	90	12.5	9.1	65	100
19	<i>Almost always</i>	95	91.6	95	5.0	6.0	60	99
20	<i>Always</i>	100	99.0	100	0.0	3.9	70	100

Table B1: Descriptive statistics for all frequency expressions (ordered by median, highest IQR and SD values marked)

Note: The median represents the central interpretation assigned by respondents. The interquartile range (IQR) contains the middle 50% of responses. Smaller IQR values indicate greater agreement among respondents, whereas larger IQR values indicate broader variation in interpretation. Standard deviation (SD), mean and mode are included for completeness.

Appendix C. *Supplementary quantitative analyses and respondent reflections*

C.1 Supplementary quantitative analyses

C.1.1 Professional background and experience

To investigate whether respondents' professional background influenced the interpretation of the frequency expressions, descriptive comparisons were carried out between language testers/raters, language teachers and respondents from other professional backgrounds. Similar comparisons were conducted for respondents with different levels of professional experience.

The analyses focused on the central tendency and spread of the numerical ratings assigned to each expression. Across all groups, the overall interpretation patterns proved highly comparable. Respondents working in language testing and assessment exhibited slightly smaller interquartile ranges for several expressions, particularly those located in the middle of the frequency continuum. However, these differences were small and no systematic differences in interpretation were observed.

These findings support the conclusion presented in [Section 3.4](#) that professional role and experience exerted only a limited influence on the interpretation of the frequency expressions.

C.1.2 Familiarity with language proficiency descriptors

Respondents also answered the question how familiar they considered themselves with language proficiency descriptors, using a five-point self-assessment scale ranging from *not at all familiar* to *very familiar*. Additional descriptive analyses were conducted to examine whether increasing familiarity was associated with greater agreement in the interpretation of the frequency expressions..

Only limited differences were observed. Although respondents reporting higher levels of familiarity occasionally exhibited slightly smaller interquartile ranges for individual expressions, no systematic pattern emerged across the complete set of expressions. Likewise, median interpretations remained highly comparable across familiarity levels.

These findings suggest that self-reported familiarity with language proficiency descriptors alone is not sufficient to explain differences in the interpretation of frequency expressions. This observation is consistent with the analyses presented in [Chapter 5](#), which indicate that practical experience in language testing and assessment appears to be more influential than familiarity itself.

C.2 Respondent reflections

The sections below summarise the responses to the two open-ended questions included at the end of the questionnaire. Respondents were first asked to indicate which expressions they found difficult to interpret (multiple responses permitted) and were subsequently invited to provide additional comments on their choices and on the survey more generally. These qualitative findings complement the quantitative analyses presented in the main report by providing additional insight into the reasoning underlying respondents' numerical ratings.

C.2.1 Expressions reported as difficult to interpret

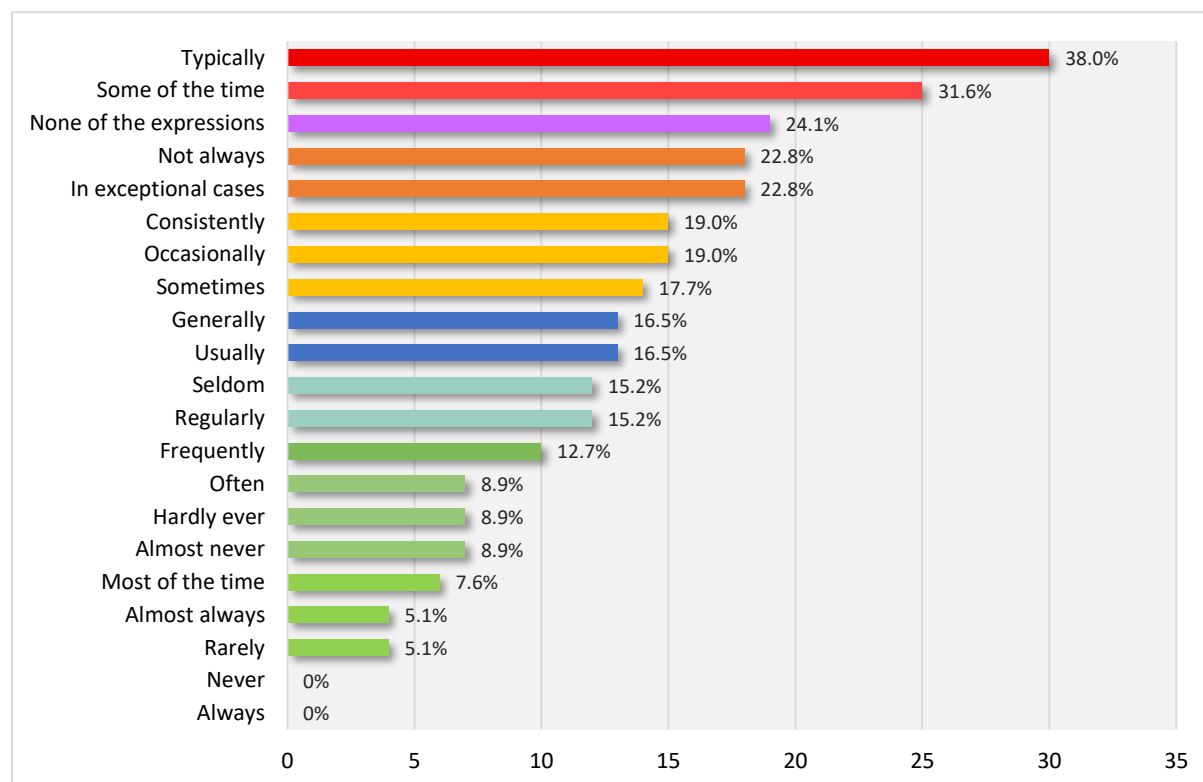


Figure C1: Expressions identified by respondents as difficult to interpret (N=79, multiple responses permitted).

Note: Respondents could select more than one expression. Consequently, the percentages shown for individual expressions do not sum to 100%. “None of the expressions” indicates respondents who did not identify any expression as difficult to interpret.

C.2.2 Respondents reporting no interpretation difficulties

As shown in Figure C1, nineteen respondents (24.1%) indicated that none of the expressions had been difficult to interpret. To examine whether these respondents constituted a distinct subgroup, their numerical ratings were compared with those of the remaining 60 participants. The comparison examined both the numerical ratings assigned to the frequency expressions and the degree of agreement within each group.

No systematic differences were observed between the two groups. Median ratings for individual expressions were either identical or differed only marginally, and there was no consistent pattern indicating greater or lesser agreement among respondents who had reported no interpretation difficulties. Likewise, Mann-Whitney U tests comparing the ratings assigned by the two groups revealed no statistically significant differences for any of the 20 frequency expressions (all $p > .05$).

These findings indicate that respondents' subjective confidence in interpreting the expressions did not necessarily correspond to objectively different interpretation patterns. In other words, respondents who reported no interpretation difficulties interpreted the expressions in much the same way as those who had identified one or more expressions as difficult.

C.2.3 Main themes emerging from respondents' comments

Although respondents highlighted different expressions, their written comments revealed a number of recurring themes. These themes provide useful context for interpreting the quantitative results presented in the main report.

Context determines interpretation

The most commonly recurring observation was that frequency expressions cannot easily be interpreted independently of the context in which they are used. Many respondents emphasised that the intended meaning of expressions such as *usually*, *typically* or *occasionally* depends on the activity, situation or descriptor in which they occur. Several respondents therefore indicated that assigning a single numerical value to an isolated expression felt artificial.

Difficulty assigning numerical values

A second recurring theme concerned the questionnaire itself. Numerous respondents commented that translating naturally occurring language into numerical percentages was inherently difficult because frequency expressions are intentionally approximate rather than precise. Several participants noted that they would rarely think about such expressions in numerical terms during normal assessment practice.

Overlap between expressions

Many respondents remarked that several expressions appeared synonymous or nearly synonymous. Groups of expressions such as *usually*, *generally*, *typically* and *regularly*, or *often* and *frequently*, were described by many as difficult to distinguish when assigning separate numerical values.

Although many respondents experienced this overlap as a challenge during the questionnaire, the finding can also be interpreted positively. Rather than indicating a weakness of the descriptor language, the responses suggest that several expressions may be used flexibly to express broadly similar meanings while still allowing natural variation in wording.

Particular difficulty with "not always"

Several respondents specifically referred to *not always* as problematic. Unlike the other expressions, it does not directly describe frequency but merely excludes complete consistency. This observation closely mirrors the quantitative findings, which showed an unusually broad distribution of numerical interpretations for this expression.

The meaning of "typically"

The single most frequently reported expression reported as difficult to interpret was *typically*. Several respondents indicated that they associated the word not only with frequency but also with what is normal, characteristic or expected in a given situation. This suggests that respondents may interpret the expression partly as a description of prototypical behaviour rather than simply as a frequency marker.

C.2.4 Illustrative respondent comments

The following quotations illustrate the principal themes identified above. Minor typographical corrections have been made where necessary, but the wording has otherwise been retained.

Context dependence

"Without context it is difficult to assign exact percentages."

"Several expressions would be interpreted differently depending on the situation."

Approximate language

"I would never normally think of these expressions as percentages."

"The exercise forced me to quantify language that I would usually interpret more intuitively."

Overlap between expressions

"Several expressions seem virtually synonymous."

"*Often* and *frequently* mean almost the same thing to me."

Special cases

"*Not always* simply means 'not 100%', so it could refer to a very wide range."

"*Typically* describes what is normal rather than how often something occurs."

