



founded 1881

July 30, 2009

Division of Dockets Management (HFA-305)  
Food and Drug Administration  
5630 Fishers Lane, Room 1061  
Rockville, MD 20852

Re: Docket No. FDA-2009-D-0181

Dear Sir or Madam,

Enclosed herein are comments on “Guidance for Industry; Label Comprehension Studies for Nonprescription Drug Products”, published as *Draft Guidance*<sup>1</sup>. The Consumer Healthcare Products Association (CHPA) is the national trade association representing the leading manufacturers and distributors of OTC medicines and dietary supplements in the United States. CHPA and its member companies have an interest and expertise in label comprehension studies and support FDA’s efforts to develop guidance for industry on this important topic. Label comprehension studies are an important tool to assess consumer understanding of information on labels of nonprescription drug products.

CHPA’s comments on the *Draft Guidance* are organized into General Comments, Detailed Comments by Section (Attachment 1), and Statistical Considerations (Attachment 2).

1. General Comments

- a. Guidelines are appropriate for label comprehension studies. There is no one template that can be applied to all studies. Many aspects of design are highly dependent on the individual application and the individual drug candidate<sup>2</sup>.
- b. CHPA recommends this guidance focus specifically on label comprehension studies. There are several instances where FDA includes references to self-selection studies. Until such time as FDA develops guidance for self-selection studies, CHPA recommends that references to how information from label comprehension studies is applied to making product use decisions be deleted. A clear statement of the use of label comprehension data should be included, e.g. to determine how effectively the label communicates information to the consumer.

---

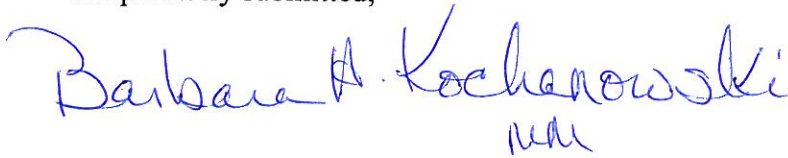
<sup>1</sup> **Federal Register**, Vol. 74, No. 83, pp 20322-20323, May 1, 2009.

<sup>2</sup> Paraphrased from comments by Dr. Eric Brass at meeting of FDA’s Nonprescription Drugs Advisory Committee meeting on September 25, 2006.

- c. The science and methodology used in label comprehension studies should be recognized as unique vs. other related fields, such as clinical trial and behavioral research. This science continues to evolve and attempts to force fit standards from clinical trial research should be discouraged.
- d. CHPA recommends the addition of a definition and discussion of the concept of “mitigation”, or other suitable term. Mitigation refers to the process of probing to understand the reasons behind a response to a close-ended question. Mitigation may lead to an initially correct or acceptable response being incorrect, or to an initially incorrect response being correct or acceptable. The protocol for the label comprehension study should outline as many of these potential scenarios as possible and how answers will ultimately be categorized. For example, mitigation would enable the sponsor and FDA to agree *a priori* when “I would ask my doctor/pharmacist” is a correct response.

CHPA and its members look forward to working with FDA to further develop this guidance and to enabling the development of well-comprehended labels for nonprescription drug products.

Respectfully submitted,



Barbara A. Kochanowski, Ph.D.  
Vice President, Regulatory Affairs

Attachments: As stated

/BK

**Attachment 1**  
**CHPA Detailed Comments on Draft Guidance for Industry on**  
**Label Comprehension (LC) Studies for Nonprescription Drug Products**

<b>Line Numbers</b>	<b>Section Title</b>
<b>I.</b>	<b>Introduction</b>
General	Recommend this guidance focus solely on label comprehension studies, as stated in lines 22-23. Text in lines 25-26 “and then apply this information when making hypothetical drug product used decisions” should be deleted. Lines 29-33 could also be deleted, as they refer to what LC “is not”.
<b>II.</b>	<b>Background</b>
71-72	Recommend stating that qualitative or quantitative pilot research (e.g. focus groups, one-on-one interviews, small base testing) is helpful to develop label prototypes (instead of or in addition to testing and re-testing, as stated).
81-82	It would be helpful to provide additional examples of “substantive labeling change”. Recommend stating that it may be appropriate to test only a specific section of a label, depending on the change.
84-85	Proprietary names are generally screened by FDA (in the case of NDAs), or can be tested with consumers to assure clarity and minimize confusion. LC is not an appropriate tool for comparing identical labels, differing only in trade name.
91-92	Recommend deleting this point. There is no rationale stated for LC testing of package inserts. Currently, package inserts are not required to be tested for comprehension. They are not visible at point of purchase and do not assist the consumer in understanding a product label at point of purchase.
<b>III.</b>	<b>Study Design and Conduct</b>
110	Clarify “when necessary” to enrich the study population. Other than low literacy, enrichment is most appropriate for self-selection studies.
118	Clarify “as close as possible”. Does this mean font size, color, graphics, etc., or does the phrase refer to word choice?
123-130	Variations in wording and information location can be studied in pilot research prior to LC, as suggested above for lines 71-72. This is preferable to testing multiple labels in LC.
<b>III.A.1</b>	<b>Primary Communication Objectives</b>
<b>III.A.2</b>	<b>Secondary Communication Objectives</b>
General	Primary communication objectives should be safety-related (higher risk), reflecting the information with potential for negative consequences if not understood. Targets should be set case by case and depend on the product and warnings under consideration. Each message should have a target, agreed <i>a priori</i> .
<b>III.B</b>	<b>Study Population</b>
183	Recommend beginning this section with a focus on the general population as the test population for LC.
185-186	Delete “whether or not individuals express interest in using the drug product” since

	this is guidance for LC and interest in using the product should not be a criteria for participation in LC.
187	Reference to enrichment should be deleted. Enrichment is appropriate for self-selection, not LC.
200	Delete “meaningful” as a descriptor of statistical analysis. Sample sizes should be determined based on targets and confidence intervals for primary communication objectives.
201-203	Regarding the concept of numeracy, as stated by FDA, there are currently no validated screening instruments for numeracy testing. Significant numbers on OTC labels generally refer to dosing. There are many ways to test a consumer’s ability to identify the right dose, and LC is not likely the best way. Numeric concepts can easily be incorporated into comprehension questions. Numeracy testing, as done today, adds 15 minutes to an already-long process, potentially leading to greater fatigue among subjects.
<b>III.C</b>	<b>Statistical Considerations and Data Analysis</b>
	See Attachment 2
<b>III.D</b>	<b>Questionnaire Design</b>
278	Typo - “illicit” should be “elicit”.
281-282	Recommend stating that every message may not need to be tested, e.g. standard OTC warning statements. Focus should be on new messages or label content that is being changed.
293	Open-ended questions are intended to verify that the consumer really understood the question, or to collect additional insights, not to “validate” close-ended questions. Reference to the concept of “mitigation” as discussed in the cover letter could be included here.
334	Pretesting is not “validation” in the statistical sense. Recommend stating “pretesting is extremely useful as a development procedure”.
336	The “bullet points” preceding this paragraph apply differently in the situation where one uses a trained interviewer vs. a self-administered test. Recommend clarifying which “bullets” apply under each circumstance.
<b>III.F.</b>	<b>Study Conduct and Location</b>
General	Customary conditions of purchase do not apply to LC testing. LC is performed to assess how people cope with materials and cognitively deal with messages, which can be done in a variety of settings, including a clinical or simulated clinical setting. Simulating conditions of purchase is appropriate for self-selection testing. Reflecting conditions of use is appropriate for self-selection and actual use testing.
<b>III.G.</b>	<b>Data Collection, Recording, and Auditing</b>
370-373	CHPA agrees every attempt should be made to pre-specify correct and incorrect answers to closed-ended questions and to follow study procedures. However, one cannot predict every possible situation. It should be recognized that additional, unexpected responses and procedural deviations may occur and that these situations need to be evaluated and their impact on the study data determined. A study is not necessarily invalid if all responses were not pre-specified or if minor procedural deviations occurred.



<b>IV</b>	<b>Final Study Report</b>
385	The statement about study subjects is confusing. CHPA's understanding is that LC should be conducted among general/representative population. The target population could be quite different.
388-390	The intent of the sentence beginning with "if possible" is not clear. If a potential subject didn't respond or agree to participate, it is not appropriate and may be impossible to encourage them to provide additional information (demographic factors, reasons for not participating). Also, these individuals will not have signed an informed consent, so it is not appropriate to collect data from them. This expectation appears to be originating from a clinical orientation (or from actual use trials), where such additional information gathering is quite usual/expected.
<b>V</b>	<b>Interpretation of Study Findings</b>
General	A discussion of mitigation (as noted in CHPA cover letter) would be appropriate in this section.

## Attachment 2

### Statistical Considerations and Data Analysis of Label Comprehension Studies

Appropriate statistical methodology is important to ensure that nonprescription products have the clearest possible labels so consumers are not confused by medicines they find in the marketplace. The *Draft Guidance* raises two major concerns:

- 1) a single approach to the conduct and analysis of label comprehension (LC) studies may discourage innovative thinking about the research problem and slow the discovery of better ways to develop labels;
- 2) it is not clear that better labels will result from the statistical methods proposed in the *Draft Guidance*; in fact, the increased sample size necessitated by the strict requirements for type I error and type I error control and inferences in subgroups is likely to lead to less iterative testing of draft labels; while ICH E9 principles are appropriate for clinical trials of drugs and biologics, CHPA does not believe they are appropriate for LC studies.

#### Label Comprehension Studies vs. Clinical Trials

LC studies differ in a number of ways from clinical trials of drugs and biologics. These differences are reflected in study conduct, statistical interpretation and, most strikingly, how study findings are used.

Clearly, LC studies are used to evaluate a *label*, not a drug. A typical LC study considers one version of the label, has minimal inclusion/exclusion criteria, recruits a study sample representing consumers who are likely to encounter the product in the marketplace, and poses no risk to study subjects since no drug is administered. Clinical trials of drugs usually study at least two treatments (including a study medication), define the study population narrowly, recruit subjects through medical practices, and potentially pose some risk to study subjects.

The consequences of a type I error are different in LC studies than in clinical trials of efficacy, safety and bioequivalence. Such an error in an efficacy study might lead to consumers being treated with an ineffective drug. In a bioequivalence study, type I error might lead to consumers being treated with a generic drug that is not equivalent to the innovator drug. Type I error in a LC study has a different implication - it means that the comprehension rate for some part of the label is lower than indicated by the study findings.

The findings of a LC study lead either to label revision and retesting (or other types of consumer research) or to use of the draft label in consumer behavior studies. If a label is found to need improvement, changes can be made fairly easily and the label tested again. This is in stark

contrast to clinical trials of drugs in that, by late phase II or phase III, a drug cannot readily be changed and tested again. If a draft label is believed to be well-comprehended, the sponsor uses it in a self-selection or actual use study. The best test of the draft label is whether consumers make correct decisions about using the product and then actually use the product correctly, that is, according to the package directions.

It is important to recognize that low comprehension rates do not predict incorrect selection or incorrect use of a nonprescription drug product once the product is marketed. LC study subjects are studied in isolation from family members, pharmacists, toll-free product information lines, and other sources of assistance and information available to consumers in the marketplace. In addition, LC study subjects are recruited regardless of their interest in the product or its indication. Subjects who are not interested in the product may not pay as much attention to the label or to the questions asked in the LC study. In summary, the LC study gives a general indication of whether the language on the label is understood by a broad cross-section of potential consumers.

### **Alternative Statistical Methodology for LC Studies**

In this section, the features of the analysis are bulleted and are followed by explanations.

- Set a threshold for each communication objective where the threshold reflects the level of potential risk faced by a consumer who does not understand the label element. These thresholds would be lower for low-risk communication objectives and higher for communication objectives that have high risk associated with lack of comprehension. Thresholds should be justified by the sponsor.

*Explanation:* Thresholds may appropriately vary across products and label elements. Thresholds should appropriately reflect the potential risk associated with failure to understand that element (lines 148-152, 157-163 and 168-171). For example, not all warnings and contraindications are of equal clinical significance either within a product or across products.

- A LC study is successful if each label element meets its target threshold. Meeting the target threshold means that a statistical test indicates that the true (unobserved) comprehension rate is at least as high as the target threshold.

*Explanation:* This definition of success requires that the lower bound on the unobserved comprehension rate for a label element meets or exceeds the target threshold. Since each individual label element must meet its threshold in order to make the conclusion that the label is understood, the experiment-wide type I error rate is controlled and there is no need for correction for multiplicity. If some communication objectives were not met, the label would be reassessed and the revised label might be tested again.

- Test the comprehension rate for each communication objective against its threshold using a one-sided test at the 5% level (lower 95% confidence limit) in the general (unenriched) sample.

*Explanation:* A one-sided test/confidence interval is appropriate since the target thresholds are one-sided. That is, the concern is that the true comprehension rate might be below the target threshold. There is no concern that the true comprehension rate is too high.

In statistical terms, the null hypothesis is that the true comprehension rate is below the target threshold and the alternative hypothesis is that the true comprehension rate is at or above the target threshold ( $H_0: \pi < P_0$  versus  $H_1: \pi \geq P_0$ , where  $\pi$  is the true comprehension rate and  $P_0$  is the target threshold). One rejects the null hypothesis and concludes that the label element is understood if the lower 95% confidence bound is at or above the target threshold.

A type I error rate of 5% is reasonable and based on precedent. Bioequivalence, for example, is typically established using two one-sided tests, each at the 5% level and is often implemented as a two-sided 90% confidence interval<sup>1</sup>. The goal of the analysis is to ensure that the ratio of mean exposure for two products is neither too high (greater than 1.25) nor too low (less than 0.80). By analogy, a label comprehension study is to ensure that the comprehension rate is not too low (less than the target threshold).

The utility of this analysis is set out in *Approved Drug Products with Therapeutic Equivalence Evaluations*, 29th edition, 2009:

The primary concern from the regulatory point of view is the protection of the patient against approval of products that are not bioequivalent. The current practice of carrying out two one-sided tests at the 0.05 level of significance ensures that there is no more than a 5% chance that a generic product that is not truly equivalent to the reference will be approved.

- Compute the comprehension rate and its one-sided lower 95% confidence bound for each communication objective for each subgroup of interest (e.g., based on literacy level, age, sex, presence of risk factors).

*Explanation:* Subgroup analyses are helpful as a diagnostic tool and are best used to explore inadequate overall comprehension rates. For example, if a label element does not meet its target threshold, comprehension rates in the subgroups might indicate whether a specific group of people did not understand it.

---

<sup>1</sup> FDA Guidance for Industry: Statistical Approaches to Establishing Bioequivalence, January 2001



The *Draft Guidance* is somewhat unclear about the role of a low-literate group in a label comprehension study. Some sections imply that low-literate consumers constitute a subgroup (lines 253-255, 395-396) while other sections imply that they are a special population, requiring a substantial sample size and compliance with the target thresholds (lines 199-201). The recommendation above treats literacy level as defining a subgroup similar to those defined by age, sex, race, and risk factors.

- The sample size for a label comprehension study should be based on the hypothesis test described above. It will incorporate the target threshold ( $P_0$ ), type I error rate  $\alpha$ , type II error rate  $\beta$ , and the anticipated observed comprehension rate for the overall (unenriched) sample ( $P_1$ ).

*Explanation:* The sample size would be chosen to provide adequate power to test the hypothesis that the true comprehension rate is less than the target threshold versus the alternative hypothesis that the true comprehension rate is at least as high as the target threshold using a one-sided 5% test/confidence interval.

The following table compares the sample size for an 80% target threshold using the methods proposed herein (a 5% type I error rate and no correction for multiplicity) with two scenarios using the methods in the draft guidance: a 2.5% type I error rate for one communication objective and a 2.5% type I error rate for 10 communications objectives (Bonferroni correction 99.75% CI).

**Sample Size Needed to Test Label Elements against an 80% Threshold**

Observed comprehension rate $P_1$	One-sided 95% CI <sup>1</sup>	One-sided 97.5% CI <sup>2</sup>	One-sided 99.75% CI <sup>3</sup>
82%	2,398	3,048	5,178
82.5%	1,520	1,934	3,281
84%	581	738	1,253
85%	365	466	787
86%	249	316	534
87.5%	158	196	331
88%	135	173	290
90%	82	107	179

1 methodology proposed in this response

2 methodology proposed in draft guidance for one communication objective

3 methodology proposed in draft guidance for 10 communication objectives

Each cell entry is the sample size needed to conclude that the true comprehension rate is at least 80% when the observed comprehension rate is as specified in the row label.

Sample size calculations are based on Fisher's exact test with 80% power.

Higher power would require larger sample sizes.

A lower threshold (<80%) would require larger sample sizes.

As shown in the table, reducing  $\alpha$  from 5% to 2.5% increases the sample size by about 27%; reducing  $\alpha$  from 5% to 2.5% and controlling for 10 comparisons using the Bonferroni correction increases the sample size by about 115%. The sample size based on the methodology in the draft guidance is substantially higher than the sample size used in recent submissions.

- In some circumstances, sample enrichment might be necessary in order to obtain a sample size sufficient to compute a precise confidence interval for the comprehension rate in a subgroup of special interest. In this case, the overall sample, without enrichment, would be used to compare comprehension rates to target thresholds while the enriched sample would be used to compute comprehension rates and confidence intervals for those subgroups.

*Explanation:* The sample size for an enriched group would be based on the precision needed for the estimate of the comprehension rate. This would be computed separately from the sample size described in the previous bullet.

- The comprehension rate for each communication objective is computed as the number of study subjects who answered the comprehension question correctly as a proportion of those who answered the question. When a self-administered questionnaire is used, the study protocol may specify sensitivity analyses to determine the effect of missing data on the estimated comprehension rates.

*Explanation:* Comprehension rates should be computed using the available data by dividing the number of subjects who correctly answered a question by the number of subjects who answered the question. Missing data are not common in interviewer-administered LC studies and those missing data reflect an interviewer's overlooking a question rather than a subject's inability to answer the question. In studies with self-administered questionnaires, missing data may reflect a subject's difficulty with the written material. In that case, sensitivity analyses may be performed to assess the impact of the missing data.